Experiments, models and methodology

By Till Grüne-Yanoff

Version: 2020-11-23

Contents

Part 1 - Methodology

1.1 Introduction – Goals of science

Welcome to Theory and Methodology of Science with Applications. This text will discuss methods: methods for producing evidence, methods for representing phenomena, methods for making inferences and for testing theory. However, the purpose is not to discuss the technical detail of these methods. Rather, the focus is on **methodology**, the systematic assessment and justification of method choice. Scientists typically must choose between alternative possible methods when doing their work. What the relevant alternatives are depend partly on what goal a scientist wants to achieve with their work. **Typical goals in science** include prediction, explanation and design.

Methodology: The systematic assessment and justification of method choice.

Typical goals of science: Prediction, explanation and design.

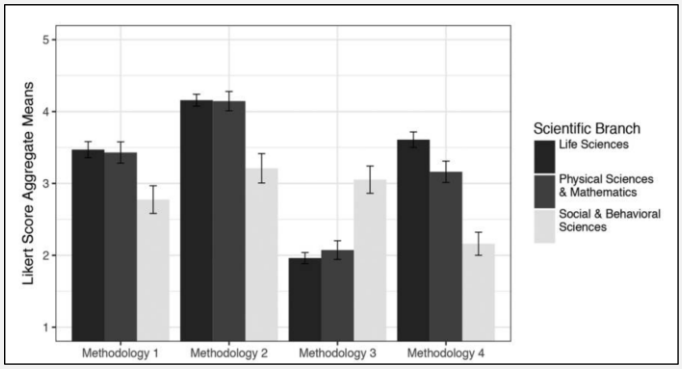
However, you cannot expect that specifying your goal immediately determines what method to choose. Instead, it will require considering the reasons why one method might serve one's goal better than another, which, in turn, might require specifying one's goal more precisely, or learning more about the methods and the context in which the methods are supposed to be applied. A couple of examples might illustrate this point. For example, your goal might be to compare how much wear different road covers can sustain under real-world traffic conditions. Which method for producing evidence should you choose? One alternative is to design a laboratory experiment. This gives you a lot of control over background variables. But the kind of use you can simulate in the lab might be quite different from the real-world traffic conditions.

Alternatively, you could run a field experiment, constructing real stretches of road with the different road covers that you want to test, and let public traffic do its damage. Here, the test conditions are realistic, but you might not be able to control all background factors to your satisfaction. These different types of experiments and their respective advantages will be discussed in much more detail later on, so do not worry if this sounds strange to you right now. The case just illustrates how a scientist must choose between different methods – in this case, different types of experiments – and that this choice requires considering the advantages and disadvantages of each of these methods for one's goal.

Take another example. Your goal might be to model a benzene molecule in order to simulate a chemical synthesis with it. Should you do this by constructing a structural formula, or by building a quantum model? The former is simpler, and this makes it easier to analyze. But the latter is more accurate. Again, understanding how important these respective advantages are will depend on how well one understands one's goal. Or take this example: you want to test the claim that street level concentration

Practice Snapshot: Multiple Conventions

A recent study asked 346 scientists from different disciplines the following four questions about their method choices. Their answers show clearly that there are different conventions of what methods to use in different scientific disciplines.



Robinson, B., Gonnerman, C., & O'Rourke, M. (2019). Experimental Philosophy of Science and Philosophical Differences across the Sciences. *Philosophy of Science*, *86*(3), 551-576.

of exhaust fumes causes respiratory diseases. Which method for hypothesis testing should you choose? Should you use a statistical test, or not? And if yes, should you perform a significance test, or employ a Bayesian approach?

These are typical method choices that scientists face. How do they decide between these alternatives? Is there a way to determine what method is rational to choose? There are three common answers to this question. The conventional answer suggests choosing the methods your teachers and your peers are using. So according to the **conventional view**, within a discipline, or maybe just within a research group, the question of method choice does not arise. Everybody just follows the established convention, just like everybody in Sweden drives their car on the right side of the road. An example of such a convention, even today, is perhaps that most scientists use statistical tests for all their hypothesis evaluations, and would never even think about writing a paper without such a statistical analysis.

Conventional view of methodology: Choices between different methods are justified by what is seen as "common practice" in the field.

Methodology 1: "Scientific research (applied or basic) must be hypothesis driven."

Methodology 2: "In my disciplinary research, I employ primarily quantitative methods."

Methodology 3: "In my disciplinary research, I employ primarily qualitative methods."

Methodology 4: "In my disciplinary research, I employ primarily experimental methods."

The problem with the conventional view is that it renders scientists less flexible in finding alternative viewpoints that could support criticism of their own discipline or group. If the use of a method contributes to an error in the research result, then it is much harder for a group that uses this method by convention to discover this mistake than for a group that does not follow a conventional method. An additional problem is that the conventional answer makes interdisciplinary work really hard. If two disciplines have different conventions about what methods to use, then bringing these disciplines together is like letting people drive on the right and on the left on the

Practice Snapshot: Funding for Denial

To get a feel how much money industry is paying for the services of sowing doubt, read this article:

Sharon Begley "Global Warming Deniers: A Well-Funded Machine". *Newsweek* August 3rd, 2007. Link

same street – that inevitably leads to mayhem, as long as people have no other reason for their choice than "but we've always done it this way!". To be able to collaborate with people from different disciplines, who follow different conventions, one must go beyond conventional methodology and instead make transparent the underlying justifications for choosing one's methods – which are understandable to people from different disciplines, and perhaps may even convince them.

An alternative answer to the question of how one should choose methods is to say one should choose the method that gives the **best results**. This this answer is on the right track, but it suffers from excessive vagueness. The idea here is that one chooses that method that serves some purpose best. But this purpose is not sufficiently specified. In particular, science often involves long-term plans, so that the final outcome from choosing one method instead of another are not known or uncertain. Take, for example, the Large Hadron Collider (LHC) near Geneva. Its construction so far has cost more than 7,5 billion euros – and the next-generation collider, now in planning stage, will cost even more (see box). What material benefits the research done here will eventually have, however, is entirely unclear – CERN's director, Bordry, is very explicit about the deep uncertainty about the potential benefits. Thus, choosing this method of investigating subatomic particles might not be justifiable by its material outcomes.

The best-results view of methodology: Choices between different methods should be determined based on what produces desired results.

Furthermore, who judges the best in "best results"? Some scientists have been making very good business in creating doubt about scientific results in the service of lobbyists and industry. Their choice of methods does not stand up to scientific scrutiny. But their choice presumably yielded results they considered best for themselves. Hence, we need to be more precise with regards to what we mean by justifying one's method choice.

Instead, I will argue for another justification of method choice, which I call the **epistemic tool account**. According to this account, one chooses that method one has most reasons to believe is the best tool for one's goals. Scientists might pursue different types of goals, for example prediction, explanation, or design. For a given method choice problem, once these goals are sufficiently specified – e.g. what is to be explained or predicted, what rate of error is permissible in the prediction, etc. – one can ask which conditions one must satisfy in order to reach these goals. For example, to explain a phenomenon Y requires identifying the cause X that produced Y. To understand this, we need to analyze what the concept "explanation" means (I will do this later in this course, so do not worry if you disagree

The epistemic tool view of methodology: Choices between different methods are determined on what one has the most reason to believe will satisfy ones epistemic goals.

Practice Snapshot: Costs of the LHC

"Cern's director for accelerators and technology, Dr Frédérick Bordry, said that he did not think that £20bn was expensive for a cutting-edge project, the cost of which would be spread among several international partners over 20 years.

He added that spending on Cern had led to many technological benefits, such as the World Wide Web and the real benefits were yet to be realised.

"When I am asked about the benefits of the Higgs Boson, I say 'bosonics'. And when they ask me what is bosonics, I say 'I don't know'. But if you imagine the discovery of the electron by JJ Thomson in 1897, he didn't know what electronics was. But you can't imagine a world now without electronics."

Cern plans even larger hadron collider for physics search, BBC News, January 15th, 2019. Link.

or don't understand yet why explanation requires identifying causes). With this analysis, we can then ask what advantages (and disadvantages) the available tools have for identifying causes in the given situation: for instance the tool of experimental methods (if experimenting is possible) or perhaps specific statistical analyses of observational data? Giving the reasons to believes that a certain method M is the most advantageous for one's goals **justifies** choosing M.

The epistemic tool account is a version of **instrumental reasoning**: it justifies choosing a method M by identifying the reasons we have to believe that M is the best tool is for a given goal in a given situation. Instrumental reasoning is widespread in everyday thinking: when choosing between colleges programs, or between burger and broccoli for dinner, people reason instrumentally, considering what goals these programs or diets serve (or do not serve). That this reasoning is so widely and successfully used is an advantage for the epistemic tool account – it indicates that this form of justification is tried and widely accepted. Note however that the outcome-oriented approach also is instrumental. The difference is that the outcome-oriented approach remained unclear on what goals it supports, while the epistemic tool account makes scientific goals specific, and it focusses on the reasons one has for choosing a method, *before* implementing it.

Additionally, in practice, there are often other considerations that are relevant for such an argument, that are not about fulfilling scientific goals. For instance, say you want to explain a phenomena Y and you know that an explanation requires accurately identifying the cause, X, that produced Y. You furthermore know that experimental methods allow for you to determine the cause of Y. If there are no other relevant considerations, then you have a good overall reason to choose the experiment as your method. However, imagine that the phenomena Y which we want to explain causes certain human patients to die. While we of course remain interested in explaining Y, there are now *legal* and *ethical* restrictions on what experiments one can perform that might cause human deaths. In this case, some reasons are favoring one method, some another. We now need to

Justification: Reasons for believing a certain **proposition** to be true.

Instrumental reasoning: Providing justification to fulfil a clearly stated goal.

Proposition: The information expressed in a statement or a claim. "Snow is white" and "Snö är vitt" contains the same proposition, but are two different statements.

Further reading

You find a detailed discussion of the epistemic tool account – where methods are described as heuristics, and methodology as guidelines for the appropriate use of such heuristics – in:

Hey, S. P. (2014). Heuristics and meta-heuristics in scientific judgement. *The British Journal for the Philosophy of Science*, *67*(2), 471-495. Link.

weigh these reasons against each other – for example, by writing a pros and cons list for each of the alternative methods in order to justify choice between them.

To conclude, this course focuses on methodology, the systematic assessment and justification of method choice. Methodology ask questions like, what relevant methods do we have available to reach a scientific goal? What reasons speak for or against choosing any of these alternatives? How should one weigh these reasons to form an overall decision? Methodology must be distinguished from describing methods on the one side and doing general philosophy of science on the other. *Descriptions of methods* concern the design and implementation of particular research strategies, and often focus on technical skills. Examples include how to program simulations, how to set up measurement instruments, how to calculate variables, or how to implement experimental design. It does not include method comparison, choice, or justification. On the other side, *general philosophy of science* asks many questions that are not directly related to method choice and its justifications either. These include questions about the nature of scientific knowledge and scientific theories, or deal with various forms of skepticism. While undoubtedly fascinating, this text only addresses such questions to the extent that they are relevant for methodological issues.

Reflecting on what you just read

- Consider a recent scientific article you have read. What was the authors' scientific goal(s)? What alternative methods do *you* think the authors had a choice between? Did the authors describe such a choice in their article?
- Describe a typical methodological convention in your field. Check with students from other disciplines whether they have the same convention or not!
- Scientists are often skeptical about philosophers' views on how scientist should do their work – arguing that scientists should know better than philosophers (who after all typically do not work in science). That seems to be a reasonable criticism, at least against those philosophers who propose fundamental and unchangeable methodological principles. Do you think that the epistemic tool account is affected by this criticism, too?

1.2 Scientific Knowledge

The main focus of this text is methodology, and rational method choice depends on the goals that scientists want to reach with these methods. Crucially, methodology, as it is presented here, does not consists of a set of *first principles* how science should be done – rather, it indicates which method is rational to choose, depending on the researchers' goals and the conditions under which they operate. The most important goals in science are prediction, explanation, and design. Being able to **predict** an event X means to know that X will occur at some time, t, in the future. Being able to **explain** why X happened rather than something else means to know what causes produced X. Finally, being able to **design** an artifact X means to know that X satisfies certain desirable functions.

Prediction: Knowing that an event will occur at a future time.

Explanation: Knowing what causes produced an event.

Design: Being able to design an artefact that satisfies certain functions.

These goals all share a common necessary ingredient: scientific knowledge. One must *know* to predict, explain, or design. And this knowledge is acquired by applying the best scientific methods. But what is knowledge? Philosophers have offered many accounts, but the answer the ancient Greek philosopher Plato gave is still considered one of the most useful. Plato argued that **knowledge** is true, justified belief. He answered the question "What does it mean that a person knows something?" The answer consists in giving necessary conditions for a person, let us call her K, to know something (that the moon is spherical, for instance), which we can call P. Then Plato's definition states that K knows P if and only if: (1) that K beliefs P to be true; (2), that K is justified in believing P to be true; and (3), that P in fact is true. These are called necessary conditions, because each of them must be satisfied for K to know P.

The classical definition of knowledge: True, justified belief.

However, even together they are not sufficient. There are cases that satisfy all these conditions, and yet most people would intuitively say that K does not, in fact, know P – search "Gettier Problem" if you are interested in such cases.

Plato's answer in the first place is just a definition, a statement what it means to know. One shouldn't just accept definitions from anybody, not even from an authority like Plato. Rather, one should ask what the arguments are for accepting such a definition. Sometimes the argument for a definition of a term is that almost everybody thinks that that term has this meaning. While such **lexical** definitions are useful to understand what others are saying, philosophers are rarely content with them. For example, even if everybody thought that the definition of a whale includes categorizing it as fish, philosophers – or scientists, for that matter – would not accept such a definition. Instead, they would argue that the definition of "whale" should not include whales being categorized as fish, even if everybody else thinks so. Such a **stipulative** definitions rest on reasons and arguments, not on common usage. Plato's definition is clearly stipulative. It is not because of this common usage that knowing requires satisfying these conditions, but because Plato and many philosophers since thought that there are good reasons for these conditions.

Lexical definition: A definition that intends to **c**apture common usage of a term.

Stipulative definition: A definition made for a specific purpose, in a specific context.

When investigating these conditions in more detail, however, it is clear that there are good reasons to adopt them as a definition of knowledge, and in particular for *scientific* knowledge. The first condition requires that for K to know P, she must **believe** it. Belief is the state of mind in which a person thinks something to be the case. Thus, this condition ties knowledge to individual state of mind. Consequently, P cannot be known without someone believing that P is true. This seems to be right – knowledge is something a person has, not something out there in the world. Scientists, when investigating the world, do not discover knowledge. Rather, they collect information, identify evidence, and craft arguments that produces knowledge by convincing themselves and others to believe that certain claims are true. Through the belief condition, it also becomes clear how knowledge affects people's behavior. People hold beliefs about what consequences their different choice options have. And based on these beliefs–combined with values, desires, and goals – they decide how to act.

Belief: A state of mind where a person considers a proposition to be true.

Knowledge has this guiding influence on actions because it is constituted by belief. But beliefs come in degrees: sometimes our belief that a certain claim is true is very strong. We mark this by saying something like, “I fully believe that”. In other cases, we hold less certain beliefs, which we mark by saying something like: “I tend to believe this”, or even, “this is hard to believe”. It seems that many of our beliefs are less than certain. There often is the possibility, and sometimes the very real probability, that they are wrong. In that case, we believe a claim up to a certain degree. The question is, however, whether such uncertain beliefs are compatible with the idea of scientific knowledge. To put it differently, does science require certainty?

The first thing to note is that Plato's definition is compatible with **uncertainty**: it only says that a person has to have a belief, not that it has to be certain. If knowledge is belief and beliefs come in degrees, then knowledge comes in degrees too. But is this how it should be? The answer is yes. Scientists often make claims that they are less than certain about– whether it concerns direct observations, measurement instruments, results of calculations, or claims about theory. These claims, if justified and true, nevertheless constitute knowledge. If you think that this sounds strange, be clear that *certainty of belief* and *truth* are two completely different things. One can have certain beliefs about a false claim. Most pre-modern people, for example, presumably believed with near certainty that the sun revolves around the Earth. And one can have uncertain beliefs about true claims. For example, if a fair coin is to be flipped 100 times, you might believe that it will fall heads 50 times. And you might later find out that indeed this is true: having tossed it 100 times, you observed it fall heads 50 times. But before you checked, you would not likely believe this with certainty. That is, you have an uncertain but true belief. In fact, acknowledging uncertainty in one's knowledge is very important in science. It is possible that even our best theories are wrong. But that does not mean what various self-declared anti-scientists want it to mean when they go on about evolution or climate change just being a theory. We have good reasons to be very confident in these theories without denying that there is a (remote) possibility that they might be wrong. We can acknowledge that we are less than certain in our knowledge without this denying its truth.

Uncertainty: A belief is uncertain if it is not absolutely justified.

Moving on, the second condition is that a belief must be **justified** to constitute knowledge. A justification is evidence or some other reason that can be put forward in defense of believing a claim to be true. Only beliefs properly justified might constitute knowledge. Many beliefs that people hold lack a justification. A lottery player presumably believes that the numbers she records on the slip will win, but she has no justification for this belief. Thus, even if the number of wins – thus her belief was true – she did not know that this number would win. Other cases are more complex. If my upstairs neighbor waters her plants and splashes my window, then if I believed that it rained because of the drops on my window, my belief is not justified and does not constitute knowledge - even if it actually rained! The observation of *those* drops (coming from my neighbor’s watering can) does not constitute evidence for the claim that it

Justification: Reasons for believing a proposition to be true.

rained. Thus my belief is true (as it actually rained), but the kind of evidence that I point to in defense of that claim is not an actually justification. So my belief that it rained is true but unjustified belief, and hence not knowledge.

Furthermore, not all relevant reasons for beliefs are justifications for scientific knowledge. If one strongly wishes something to be true, the happiness one feels if one believes that it is true might be a strong reason why someone believes it. But such wishful thinking does not constitute a justification of scientific knowledge, nor do religious convictions. One can, of course, have these beliefs, but they do not constitute scientific knowledge.

A belief is justified if evidence or some other reason can be put forward in its defense. But do we actually have to have this evidence or other reason "in hand", so to say? Most Swedes, for example, know that Gustav Vasa fled on ski through Dalarna from men loyal to King Christian. Yet what kind of justification can they show for this belief? Of course, no living person today has witnessed this event, so they instead must rely on documents or reports about it. But most Swedes learn about this event in elementary school – and while they remember the story, they most likely don't remember where they learned it. So they cannot point to the actual reason that made them form this belief – and that might seem to imply that their belief is not justified. The position that I defend here, called "weak internalism of justification", contradicts this impression. Someone's belief is justified only if one can become aware by reflection of what the belief's justifiers are. This reflection might involve considerable time and effort (for example, researching what history textbook you had in fourth grade, etc.); the point is that one is justified even if one does not have evidence or reasons "in hand" – it is enough that one could get hold of them, if one invested sufficient effort. Thus, when challenged on a certain claim, one must in principle be able to produce the justifiers for them. It will not do to insist that these justifiers are to be found somewhere, even though one cannot, even with one's best efforts, locate them anymore.

Practice Snapshot: From uncertainty to falsity?

Here's an example of an argumentative strategy that implicitly assumes that knowledge must be certain:

"How did life originate? Evolutionist Professor Paul Davies admitted, 'Nobody knows how a mixture of lifeless chemicals spontaneously organized themselves into the first living cell.'[1] Andrew Knoll, professor of biology, Harvard, said, 'we don't really know how life originated on this planet'.[2] A minimal cell needs several hundred proteins. Even if every atom in the universe were an experiment with all the correct amino acids present for every possible molecular vibration in the supposed evolutionary age of the universe, not even one average-sized functional protein would form. So how did life with hundreds of proteins originate just by chemistry without intelligent design?"

Source: https://creation.com/15-questions-for-evolutionists

For a discussion of argumentative strategies of climate deniers – including arguments starting from the assumptions that knowledge must be certain – see this article and short video introduction: https://iopscience.iop.org/article/10.1088/1748-9326/aaa49f

Scientists are expected to publicly defend their scientific beliefs, and they are expected to make public their justifications for these beliefs. Numerous scientific practices and institutions are designed towards this public justification goal, be it public thesis defenses, seminar and conference presentations, the peer review journal system, the tenure track system, insistence on data transparency, or attempts to reproduce experimental results. All these designs share the basic idea that although justification is crucial for science, few claims can be justified conclusively so that no doubt remains. Instead, science encourages us to continue scrutiny of justifications and allows for any reasonable challenge to be heard and responded to. It is worth reminding ourselves that such a system did not always exist. Just 400 years ago, Giordano Bruno was burned on the stake for not bowing to the authority of the church in astronomical questions. Galileo only avoided that same fate because he did bow. It is only rather recently that the exchange of criticism and the respect for reasons and arguments has replaced such authority-driven systems, and it is still not universal.

Further reading

If you are interested in reading more on why the heliocentric model devised by Copernicus was no more justified by 16th and 17th century data than Ptolemy's system, read:

Gingerich, O. (2011). Galileo, the Impact of the Telescope, and the Birth of Modern Astronomy1. *Proceedings of the American Philosophical Society*, *155*(2), 134-144.

The third and final condition of the classical definition of knowledge is that knowledge is **true** justified belief. How to characterize truth is one of the big philosophical problems. Let us instead use a very simple conception: true statements describe (part of) how things really are. As argued before, belief and truth are two different things. Even if I believe with certainty that P is true, P might still not be true. Perhaps more interesting, *even* if I am justified in believing that P is true, P might still not be true. This is most obvious when looking at historical examples. Astronomers before the invention of the telescope believed that the earth was in the center of the universe and that all heavenly bodies moved around it. Moreover, they were justified in this belief. The observations they could make with their technology were not sufficiently precise to reveal planetary movements that would have contradicted that belief. Based on the available data at the time, the geocentric model was justified. Pre-telescope astronomers thus held false justified beliefs.

Truth: A proposition is true if it describes how things really are.

This example shows the importance of adding truth as a third condition for knowledge. We do not want to say that those astronomers *knew* the positions of the earth in the solar system, but we do not want to deny that their beliefs at the time were justified. The problem is that today, we might not be in a much better position than those pre-telescope astronomers were. Some of our scientific beliefs might be justified and nevertheless false. We therefore might not have scientific knowledge. But how do we find out? We cannot access truth directly. All we can do is to justify our beliefs and hope for the best.

Empiricism: Knowledge arises from evidence gathered via sense experience.

One possible solution is to start from the view that we might at least have access to the truth of some of our beliefs, and then argue from their truth to the truth of others. One such approach that many scientists agree on is **empiricism**, along with the assumption that we can tell whether some

observational reports are either true or false. Such reports are more about us, the observers, than about the world. And because most humans share the same observation abilities, they might be able to determine the truth even of other people's observational reports. But this leaves the possibility of metaphysical skepticism. Even if we can tell the truth of observational reports, the skeptic might say, how do we ever learn about the world itself? Science, after all, talks about trees and cars and electrons, and forces and the regularities between them – not about observations of cars or cars, or electrons, etc. How do we ever determine the truth of statements about these things from the purported truth of observations about these things?

So, on a rather fundamental level, our discussion of knowledge has led us to the question what science and scientific knowledge is really about. Philosophers have different opinions on this question. **Instrumentalists** argue that creating a theory is just to rearrange and order observational reports. The theories might be deemed useful, but they are never strictly speaking true or false. **Realists** in contrast argue that theories about trees, cars, but in particular about not directly observable objects like electrons or forces, are either true or false. This realist position is appealing to many, including many scientists. This so-called *scientific realism debate* is an important part of philosophy of science, but it has few if any methodological consequences – it therefore is enough to barely sketch it here, and move on!

Observational report: A statement about sense experience.

Instrumentalism: A scientific theory is not strictly speaking true or false, and the entities it proposes are conceptual tools, rather than something that exist.

Realism: A scientific theory is true or false, and the entities it proposes either exist or do not exist.

Part 2 - Scientific inference

2.1 Scientific inference rules

I have given you examples of methods and method choice. These examples were largely related to specific observations, such as observations in an experiment or when manipulating a model. But science is not generally interested in the particular: scientists want to generalize. They want to say something beyond the particulars that they observe and predict new phenomenon or explain phenomena that they did not directly investigate. In order to do so, they need to make **inferences**: starting from something they know and going to something else that they do not know yet. This is what we call an inference – an act or a process of reaching a conclusion from some known facts or evidence. Think of this as a simple relation between what we call a **premise** and a **conclusion**.

There are many different kinds of inferences, and there are even more specific inferences or inference rules. When we have a lot of different inference rules that we could use, we have to make a choice, and we come back to the general theme of method choice. We have a menu of methods to choose from, and we have to justify which of these methods we are choosing. So let us go and have a look and see some of the examples of inferences. A very common one is a **generalization**. In a generalization, one infers from a sample to a general claim, for example: "since these one hundred and fifty blackbirds have been observed to be nest-making, we infer that *all* blackbirds are nest making". Another type is this one: "since the last five rockets reached a top speed of 12 km/s, the next one will also reach a top speed of 12 km/ s". We are projecting what we have observed

Inference: An act or a process of reaching a conclusion from a set of **premises**, which can express, for instance, known facts or evidence.

Premise: a statement in an argument that justifies a conclusion.

Conclusion: A statement that follows logically from premises.

Generalization: Inductive inference from a sample to a general conclusion.

onto future cases that we consider to be of the same type. This is another form of inference: **projection**. With this kind of inference, we are still basing it on a *finite* number of observations, but contrary to the first case, we are not drawing a general conclusion from it (to a potentially infinite number of cases) but only a conclusion about another small finite number of objects.

Projection: Inductive inference from past samples to future samples.

Now, these two types are somewhat similar. They all start out with observing a finite number of objects and then draw conclusions about objects of this type above and beyond the original observed ones. These are called **inductive inferences**. They are commonly contrasted with a different type of inferences, where you start with a general assumption and infer something particular: called **deductive inferences**. Such inferences are common in mathematics, but also in physics or economics.

Further reading

For further discussion of different types of inductive inferences, of the need to justify the choice of inference rules, and it dependence on facts, read:

Norton, J. D. (2003). A material theory of induction. *Philosophy of Science*, *70*(4), 647-670.

Inductive inference: In an inductive inference, the premises support the conclusion but does not guarantee its truth.

Deductive inference: In a valid deductive inference, true premises necessitate the truth of the conclusion.

A common characteristic of this other category of inferences is that they often involve a **conditional claim**. This is not as complicated as it might sound, if you are saying to a friend, "If John bought ice cream, James bought strawberries", you are technically making a conditional claim. If it is true that John bought ice cream, you can infer that James bought strawberries (but God knows what happened if John did not buy ice cream!). More formally, a conditional claim is saying that *if* a number of assumptions are true, *then* the conclusion is true. If one then finds (or assumes) that the assumptions are true, one can infer the conclusion is true. That is a type of deductive inference called **modus ponens**. It is a standard type of inference in propositional logic. Here is another type of deductive inference, a type, which will be very important when we later on discuss *falsification*. We say, well, we start out with this conditional: "If the hypothesis is true, then certain consequences must be true". And then we observe that the consequence is false. And thus, we infer that the hypothesis must be false. The conditional claim could be "If it rains on this Saturday, then the music festival will be canceled". If this claim is true, and it is also the case that the music festival was not canceled, you can draw the conclusion (i.e. infer) that it did not rain on that Saturday. That is what we call a **modus tollens**.

Conditional claim: A claim involving the logical operator "if", for instance of the form "If A then B".

Modus ponens: A deductive inference of the form: *If A then B*, *A*, therefore *B*.

Modus tollens: A deductive inference of the form: *If A then B*, *not B*, therefore *not A*.

I have here given you a number of examples of types of inference rules. The first two types of inferences are inductive. They are inductive in the sense that they *amplify knowledge*. And here, **amplifications** – strengthening, boosting, or other similar terms – mean that we extend the knowledge that we have in the premise to something beyond itself. We are observing a certain number of individuals (or events, or phenomena etc.), and then we are making claims about new individuals that were not actually in the original premises. Because we are amplifying in this way, we are also running the risk of making an error, of making the wrong kind of conclusion. It cannot be excluded in using any of the inferences that we might be wrong: even if it is correct that all observed blackbirds are making nests, the conclusion that *all* blackbirds are nest making might still

Amplicative: Inferences that go beyond what is stated in the premises – in particular, inductive inferences are amplicative.

be incorrect. The methodological question here is thus: which inductive inference rules are justified in this specific situation – and why? When are we correct in projecting from our sample to the population? This is what a lot of the discussion regarding what good inference rules are, versus what bad inference rules are, is about.

In contrast to that, deductive inferences are not amplifying. Instead, they are **explicative**, or they *explicate knowledge*. The idea is that all the information is already contained the premise. When you are making a deductive inference, you are just unpacking the information that is already contained in the premise. Think of the modus ponens example where James bought strawberries if John bought ice cream, and John did buy ice cream. Some of you might have said, "Well, this is very trivial". The conclusion can practically be found in the premises. This trivialness, in a way, is the strength, as we will see soon. And in any way, it is certainly the character of deductive inference. Namely, it only unpacks what is already at least implicit in the premises. And the good thing about this is that if you are explicating the premise, and you are doing that with a rule that is said to be *valid*, then you cannot go wrong. Deductive inferences are said to be **truth-preserving**: they ensure that the conclusion is always true if the premises are true. And that distinguishes them from inductive inferences, which are **fallible**: the conclusion might be false even if the premises are true.

Explicative: Inferences that do not go beyond what is stated (implicitly) in the premises – in particular, deductive inferences are explicative.

Truth preservation: The conclusion must be true if the premises are true, see **deductive inference**.

Fallible: The conclusion can be false even if premises are true.

2.2 Hume's problem and justifying inference rules

In the last section, I presented two *types* of inference rules and categorized them as either inductive or deductive. For each type, there are many *particular* inference rules. Take, for example, generalization, which is an inductive inference rule. One particular instance of a generalizing inference rule is rule A: "Whenever you have observed at least 150 objects of kind X to have property R, then conclude that *all* objects of that kind have property R". For example, you might ask me, why do you believe that all blackbirds are nest making? And I might answer, I observed 150 blackbirds and, using inference rule A, I am justified in believing that all blackbirds are nest making. This is probably not a very good rule – it is unclear why one should sample 150 objects, what the role of kind X is in this rule, and why one can draw a universal conclusion.

Although inference rules that scientists actually use look a bit more complicated, they have the same structure. Rule A might be very simplistic, but scientists do use something akin to rule B: "Let Q be evidence against hypothesis H. Whenever the probability of observing Q, given that H is true – this is the so-called p-value – is smaller than the significance level of 0.05, then reject H". Don't mind the details of this rule at this point – I will discuss this later in the statistics section.

However, there are there are many other particular inference rules. In fact, there is a continuum of such rules. An alternative to our simple rule A is to insist on having made at least 200 observations (or 250, or 300 etcetera), or an alternative that this holds for properties of kind Y. An alternative to rule B is to use significance levels of 0.01 or 0.1 or, in fact, any other

number between 0 and 1. Now recall what these rules are used for. They are used to justify inferences. They are used for the purposes of justifying *why* they accept or reject certain hypotheses. But B is just *one* rule on a continuum of rules. So why would scientists use B instead of some of the infinitely many alternatives? In other words, how is the choice of an inference rule justified?

Particular inference rules thus justify conclusions *only if they themselves are justified*. Be careful to distinguish these two issues – first, that inference rules justify conclusions; and second, that the choice of a particular inference rule itself must be justified. I will now focus on the justification of inference rules, in particular inductive inference rules. The justification of deductive inference rules is also sometimes discussed, notably by Charles Dodgson (who, besides being an eminent scholar, also wrote under the pen name Lewis Carroll), but I will instead focus on the skeptical argument against induction. The most famous version of this argument was developed by the Scottish philosopher David Hume. Hume published his so-called problem of induction in 1748, but it remains as fresh and unresolved as it was then. You find the structure of his argument in the box below, but I will discuss it in detail here.

Further reading

You can read Carroll's allegorical dialogue discussing problems for deduction here:

Carrol, L (1895). *What the Tortoise Said to Achilles*. Mind. Link.

Hume's argument has five steps: (1) It starts by assuming that there are only two kinds of inferences, inductive and deductive. He uses these categories in pretty much the same way as I presented them earlier, so there is nothing surprising here. In his second step (2), Hume assumes that to justify any inductive inference rule **I**, such a rule must itself be inferred from some premise. Again, this seems pretty plausible, and it is often claimed that all justificatory arguments take on the form of inference. Hume then assumes (3) that such a justification of **I** cannot proceed deductively. Making a deductive argument would require a premise like "If an inductive inference rule has worked in the past, it will work in the future". But this premise is quite clearly not true, and Hume argues that there is no other deductive type of inference rule that would work. Or, as Hume himself puts it, just because an inference rule has yielded true conclusions in the past does not necessarily imply that it will yield true conclusions in the future. Again, this seems pretty plausible as many of us have the experience of an inference rule that fails even though it performed well in the past.

1. Every inference is either an induction or a deduction
2. To justify an inductive inference rule **I**, this rule itself has to be inferred from some premises
3. **I** cannot be inferred deductively, because there are no *necessary* connection between past and future inferences
4. Thus, **I** must be inferred *inductively*
5. When inferring **I** inductively, we must appeal to another (inductive) inference rule **J** to justify this induction. But that raises the issue of how to justify **J**, which would require appealing to another inference rule **K**, ….. [*infinite regress*]

Consequently, no inductive inference rule can be justified

From assumptions one and three, Hume then concludes (4) that the justification of **I** must proceed inductively. This is just a modus ponens conclusion, so no doubt here, but it is also pretty plausible as an intermediate result. Our only chance to argue that an inductive inference rule is a good one is by showing that it has worked well in the past *sufficiently often* so that we inductively infer that we can be quite confident, though of course never sure, that it will work well in the future. But now comes the decisive blow (5). If the justification of **I** must proceed inductively, then there must be another rule **J** according to which this inductive inference is performed. And then that rule **J** also requires a justification for which we could need another inference rule **K**, and so on, leaving us with a never-ending quest for more justification. Philosophers call such a runaway chain an **infinite regress**. Ending in such an infinite regress demolishes any hope of obtaining a justification for what one wanted to justify. Hence, Hume concludes that no inductive inference rule can be justified. This is a very powerful argument, which seems to demolish our hope to justify inductive inferences. But science uses inductive inferences all the time – does not this then mean that scientists engage in practices that are not justified?

Infinite regress: A never ending chain of propositions being justified by other propositions which in turn are justified by other propositions and so on.

Earlier, I argued that basing one's claims on reasons is one of the hallmarks of science. In the light of Hume's argument, does this characteristic then disappear? Is science irrational because it lacks justifying reasons? Let us consider this claim about irrationality and science in more detail. Scientists make inductive inferences. But by Hume's argument, any attempt to justify inductive inferences ends in an infinite regress. Thus, scientists employ unjustified methods and science is therefore irrational – quite an unsavory conclusion indeed. If we do not want to accept this, then we need to dispute either of the two premises. One possibility is to deny that scientists need to employ inductive inferences altogether. That is exactly what the 20th century philosopher Karl Popper argued. I will come back to that later.

Another possibility is to deny the *impact* of Hume's argument, rather than the argument itself. Remember, Hume's argument as such has not been disproven. The infinite regress is here to stay. However, perhaps there is an interpretation of justification, which is not completely destroyed by such an infinite regress. So, what is the nature of justification? What is that which offers justification? Similar to the construction of buildings, justification, as envisioned by Hume and others, seeks an **ultimate foundation**, a set of basic claims, onto which all the other claims can be built on, or inferred from. Hume's argument denies an ultimate foundation for inductive inferences: at any level, we can always dig deeper into the ground, searching for a firmer foundation. But we never hit bedrock – that is the infinite regress. If there is no proper foundation, then one can neither erect a building, nor can one ground justification for any claim.

Foundationalism: Propositions are justified by being inferred from foundational premises which do not need additional justification, for instance necessarily true premises.

But perhaps we need no ultimate foundation to justify induction? That way, we could sidestep Hume's argument, without needing to deny its validity. But this would require a different idea of what justification is –

one that is not affected by Hume's argument. Luckily, in philosophy, the foundationalist view is not the only account of the nature of justification.

Coherentism: Propositions are justified by being compatible with a coherent set of propositions, where each proposition in the set is compatible with every other proposition in the set.

An alternative to foundationalism is so-called **coherentism**. Its basic idea is that the justification for a claim increases the better it fits into a coherent system with other beliefs one holds, or claims one accepts. To illustrate this position, the Austrian philosopher Otto Neurath used the metaphor of the ship at sea. Scientists are like sailors who, out on the open sea, must reconstruct their ship to keep it afloat but who are never able to return to dry dock to lay a new foundation. Coherentism might not suffer as much from the impact of Hume's argument as foundationalism since inductive-inference rules might not function as foundations of inductive practices. Rather, they might serve as abstract descriptions of inductive practices, and they might serve as tools that can connect inductive practices with each other and with other, more-abstract, arguments and intuitions. Then both the inductive practices and their corresponding rules are justified if they cohere with each other, i.e. if they form a coherent system. This holds for claims held by one scientist, as well as for claims made by different scientists.

We thus have a way of dealing with Hume's problem, but only by substantially changing our view of what justification is. If we accept this new view, then formulating inductive-inference rules, as, for example, statistics does, is not a deeper or more fundamental endeavor than conducting everyday science. Rather, it is a strategy for increasing coherence between practices and rules and thereby contributing to the justification of both.

2.3 The Hypothetico-Deductive Method

We have discussed different types of inference rules, particularly deduction and induction, and we have dealt with Hume's problem of induction. One conclusion from the discussion so far is that if we want to counter Hume's argument, we better pay close attention to the actual practices of induction that scientists use. We cannot say that the actual practices are only justified if they cohere with the foundational principles. Rather, as I have argued, from the coherentist perspective it is the inferential practices themselves that we should be looking at, and consider which of them are justified in which situation. Therefore, I will discuss one such practice in particular, namely the Hypothetico-Deductive method. The Hypothetico-Deductive (HD) method is widespread family of inductive inferences in the sciences. It features both deductive and inductive steps – but because the inductive steps introduce fallibility and thus prevent truth-preservation, it must be, according to the above distinction, counted as a type of *inductive* inference rule. Here are the five steps of the HD method:

1. Formulate a hypothesis H
2. Deduce observable consequences $\{C_i\}$ from H.
3. Test whether $\{C_i\}$ is true or not.
4. If $\{C_i\}$ is false, infer that H is false.
5. If $\{C_i\}$ is true, increase confidence in H

To summarize, the HD methods starts (1) by proposing a novel hypothesis H that might be true. From this hypotheses, (2) one or more observable consequences $\{C_i\}$ are deduced. Through some empirical practice (for example, an experiment), some of these consequences are (3) tested: it is checked whether these consequences of the hypothesis are true or not. The result of this will either increase or decrease one's confidence in the hypothesis. If a consequence of H turns out to be false, (4) one reduces one's confidence in H, or even rejects H altogether. This latter, stronger, inference is called **falsification**. If (some of) the consequences turn out to be true, (5) one increases one's confidence in H, called **confirmation**.

But let look at step 1 in more detail. What is a **hypothesis**? The Hypothetico-deductive method starts with the formulation of a hypothesis. Does that mean that we can formulate anything? No, it does not. There are requirements for what a hypothesis is, and certainly for what makes a good hypothesis. The first thing to note is that a hypothesis must be either true or false. That excludes a lot of statements that you can make, or questions that you can raise. In particular, it *excludes* research questions. There is a difference between asking: "what are the consequences of experiment X" and proposing: "the consequence of experiment X will be Y". A question cannot be either true or false. Questions do not have a truth-value. A proposition, a statement, however, does have a truth-value. Only those can be proposed as hypotheses.

The second requirement is that a hypothesis should not be a **tautology**. Tautologies are claims that are *necessarily* true or necessarily false. For example, in English, the term 'bachelor' is often defined as an unmarried man. Therefore, a claim that bachelors are unmarried is necessarily true, i.e. there is no evidence that can render it false (or increase our confidence in it, because what it is to be a bachelor is to be an unmarried man by definition! Inductive inference do not affect our confidence in tautologies – they are either true by definition or inconsistent (like the claim "the sum of angles in a square is 180°"). Therefore, hypotheses should not be tautological.

A third requirement is that a good hypothesis should either have some amount of generality and/or be about something that is not directly observable. Many hypotheses are not about some particular thing, but about things of that type generally speaking. If a hypothesis nevertheless is about something particular, that particular something should be something not directly observable (I will discuss the notion of direct observability in the next section). Consider the hypothesis H: "Object X smells distinctly of rotten eggs". This is a hypothesis about a particular – object X. That is already quite uncommon in science – hypotheses are

Falsification: Rejecting a hypothesis as a result of an empirical test.

Confirmation: Increasing the confidence in a hypothesis as a result of an empirical test.

Hypothesis: A proposition that can be true or false but is not necessarily true or false, and that preferably either has some generality or is about something not directly observable.

Tautology: A proposition which is necessarily true or false.

typically formulated about types, not particulars. For example, "Hydrogen Sulfide smells distinctly of rotten eggs" says that *every* object that is Hydrogen Sulfide has this odor. But H only focusses on a particular object X. Furthermore, H has really only one empirical consequence – itself. It reports a directly observable property – furthermore one that can arguably only be tested through direct sense experience, namely by smelling it. To allow a hypothesis like H into the HD method makes the whole exercise trivial. Of course, hypotheses about particulars that are not directly observable are perfectly ok – as for example the claim, "This object has an electrical resistance of X Ω". But hypotheses should not consist of claims about directly observable particulars.

Back to the Hypothetico-Deductive method, step 2. Once we have formulated a good hypothesis, we set out to deduce some observable consequence from the hypothesis that we can test. The consequence should be about something **directly observable**, or it should be about something observable with the help of some instrument, some aid. Alternatively, the observation might involve some **operationalization**. In any case, we need to have deduce a clear, empirically observable consequence that we can test. It would be pointless to derive consequences that cannot be observed, and therefore, cannot be tested, since we test hypotheses by means of testing observable consequences in the Hypothetico-Deductive method.

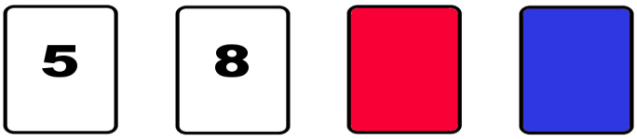
Direct observation: Observations of objects and properties that are accessible through the use of human senses.

Operationalization: A way to measure something which cannot be directly observed or that cannot be observed directly with sufficient precision, by connecting this feature with something causally connected to something that can be observed directly.

The deduction of the observable consequence must, of course, be valid, and furthermore the consequence must be *relevant* to the hypothesis and the research question. Formulating a general criterion of relevance for the *confirmation* of a hypothesis is difficult, although most people find it to be reasonably easy in practice when looking at particular hypotheses. Sometimes one can derive consequences from a hypothesis, but observing them is insufficient for increasing confidence in the particular hypothesis. These are consequences that we would expect to turn out true regardless of the truth or falsity of the hypothesis at hand. For example, from the hypothesis "Mammary cancer in mice is caused by X" one can correctly deduce the observable consequence "Mice have mammary glands". This is a valid observable consequence derived from the hypothesis. But observing such mammary glands in mice does of course not confirm the hypothesis about cancer, which was what we was interested in, and is thus not relevant.

The above relevance problem does not arise for falsification, as the falsity of *any* valid consequence from a hypothesis by Modus Tollens implies the falsity of the hypothesis. If mice do not have mammary glands, they cannot have mammary cancer. But there is a related problem, namely knowing what the relevant consequences are from a hypothesis are. To

falsify a hypothesis one must check *all* relevant hypothesis. Consider the following scenario. Four cards are presented to you like this:



A **B** **C** **D**

Now consider this hypothesis: "If a card shows an even number on one face, then its opposite face is red." What consequences do you need to observe, i.e. which cards do you need to turn, in order to check whether this hypothesis is false or not? Try to give an answer before you continue!

The hypothesis is formulated as a conditional statement; it says that *if* the number is even, *then* it will be red. In order to show it to be false, we therefore have to figure out when the statement is false. The only circumstance under which the statement is false is when the number is even but the color is not red. Therefore, you need to turn over those cards that show an even number and those cards that are not red. If a card is red, then whatever number is on the other side is fine. But if a card is not red, then we need to make sure that it does not have an even number, because if it did, that would show that the hypothesis is false.

The card example shows that it can often be tricky to figure out what the relevant observable consequences are, that you need to deduce from your hypothesis in order to be able to obtain test result data that allows us to say something about our hypothesis.

Once we have performed our deduction, derived our observable consequence, we proceed to testing it. To test it we make a direct observation, a direct aided observation, or an indirect observation to determine whether the consequence is true or not. This will be discussed more in the lectures on measurements and on experiments, when we look into scientific empirical practices.

A test of the observable consequence can, ideally speaking, yield two outcomes. Either the observable consequence is shown to be true (i.e. the observation that we had derived, or expected, did occur), or it is shown to be false (i.e. the observation that we had derived, or expected, did not occur). In the first case, we increase our confidence in the hypothesis (I will discuss confirmation in section 2.5) In the second case, it is usually argued that we can infer the falsity of the hypothesis. This is called falsification, and I will discuss this in the next section.

Remember to distinguish **falsification** in the context of hypothesis testing, where it means to prove a hypothesis false, from **falsification** in the context of ethics, where it means the unethical behaviour of changing data to suit your opinion or target.

2.4 Falsification

Many authors claim that there is a fundamental difference between confirmation and falsification in the HD method. In the final step of the hypothetico-deductive method, falsification of the hypothesis is inferred from the falsity of even a single consequence. In cases of confirmation, in contrast, we do not strictly infer the truth of a hypothesis even in the face of a very large set of true consequences. Rather, we only infer that one might be more confident in the hypothesis if suitably many consequences of this hypothesis are found to be true. This is often called the **asymmetry** between falsification and confirmation. It is widely noted amongst practicing scientists.

Asymmetry between falsification and confirmation: No amount of confirming observations can deductively confirm the hypothesis, but one falsifying observation can deductively falsify the hypothesis.

Albert Einstein succinctly described the asymmetry as follows: “no amount of experimentation can ever prove me right. A single experiment can prove me wrong”. One must read Einstein carefully here. By proof, he meant valid deductive inference. Thus, he did not deny the possibility of confirmation. But in the case of confirmation, we make an inductive inference, not a deductive one. Falsification, in contrast, is an example of a valid deductive inference.

This asymmetry between confirmation and falsification also forms the basis of Karl Popper's **falsificationism**. Popper's method closely follows the structure of the hypothetico-deductive method. According to Popper, scientists (should) conjecture **falsifiable** hypotheses. They (should) seek to falsify these hypotheses by checking whether any of their implications conflict with empirical data. They reject any hypothesis thus falsified.

Falsificationism: The view that science should proceed only through valid falsification, and never use confirmation.

Falsifiable: A hypothesis is falsifiable if it is possible to show that it is false, even if it has not yet been shown to be false.

Note the distinction between a hypothesis being falsifiable and a hypothesis being falsified. Falsifiability is a property of a hypothesis, or more broadly, of a theory. Falsifiability in this sense means that the hypothesis or theory has implications whose truth or falsity can be determined by means of observations. Some theories have no relevant empirical implications whatsoever. Popper claimed that theories in, for example, astrology, 20th-century Marxism, and Freudian psychoanalysis, were not falsifiable. He argued that the falsifiability of a theory **demarcates** it as scientific. A theory that is not falsifiable is not scientific according to Popper. Amongst hypotheses that are falsifiable, some might be more falsifiable than others because they have more empirically observable implications. Popper thought that a more falsifiable hypothesis is better than a less falsifiable one.

Demarcation of science: Distinguishing science from non-science by providing criteria for counting something as science.

Falsification, in contrast, is the event of observing that an implication of a hypothesis is not true, from which it is inferred that the corresponding hypothesis is not true. Falsifiable hypotheses need not be falsified, but hypotheses can only be falsified if they are falsifiable. Importantly, Popper restricted his testing method to the falsification of hypotheses. In his hypothesis testing method, confirmation plays no role at all. He was very explicit in his belief that one should never infer the truth of a hypothesis from observing its implications, not even increase one's confidence in the truth of a hypothesis. All that falsificationism allows is to record that hypotheses, and by extension theories consisting of such

hypotheses, have *so far not been falsified*. This is really quite a radical departure from scientific practice. Most scientists talk and act in ways that indicates that they have differencing degrees of confidence in various claims. Popper suggests that this is all bad method. But why does he insist on such an austere program?

Popper's reason for this strict aversion against hypothesis confirmation was the hope of avoiding Hume's problem of induction. Recall that Hume's argument seriously damaged any foundationalist justification of inductive inferences. Popper aimed to circumvent this challenge by proposing that science could proceed *without induction* altogether, and indeed, a scientist who strictly follows Popper's falsificationist program can do so. In the HD method, testable consequences C of a hypothesis H are *deductively* inferred modus ponens: If H, then C; H; hence C. The falsity of the hypothesis is inferred from the falsity of the implication in accordance with the deductive inference rule modus tollens – if not C, then not H. By arguing that science does and should proceed by deductive inferences *only*, Popper concluded that the lacking justification for inductive inferences is not so damaging for science after all.

However, one might worry that Popper's cure for Hume's problem have worse consequences than the disease it is intended for. This is because Poppers falsificationist program forces science into a very austere program that scientists do not actually follow in practice, and, which they arguably should not satisfy. Two arguments in particular have been presented against Popper's version of falsificationism.

The first worry is that Popper's deductive program does not allow scientists to separate between non-falsified hypotheses based on our level of confidence in them. At any given time, there are many non-falsified hypotheses. Some of these will contradict each other. Some will be connected to several observations made by scientists. Some will have been used in many explanations. According to Popper, scientists must nevertheless suspend their judgement about such hypotheses being more likely to be true than other, competing hypotheses: all that they can say is that some hypotheses are false, nothing more. That however, does not accord with how scientists act in practice. On the contrary, scientists not seldom make precisely such judgments, and it is not obvious that they shouldn't. Popper himself has proposed a measure of **corroboration** which distinguishes among non-falsified hypotheses according to the number of unsuccessful falsification attempts that they have been exposed to. I find this to be an interesting proposal, but it is unclear how this can be part of a purely deductive program. So, as a purely deductive program, Popper's falsificationist account is too restrictive to fit scientific practices.

Corroboration: A hypothesis is corroborated if it has withstood multiple falsification attempts.

A second worry concerns the applicability of modus tollens for the purposes of rejecting a hypothesis. In scientific practice, hypotheses rarely have immediately observable consequences. Instead, they often require the correct application of measurement instruments or experimental designs. Take, for example, a simple hypothesis such as: this liquid contains three chemical substances. What are the observable

consequences of such a hypothesis? Perhaps one might use distillation, observing that three fractions condensate at different temperatures. Or one might use column chromatography, observing that compounds move through the column at different rates, allowing them to be separated into three fractions. In order to deduce such observable consequences from the hypothesis, we need to make additional assumptions about the techniques involved. Let's assume that the hypothesis is true. Even so, one can expect to observe three fractions condensing only if the distillation operates properly. Similarly, one can expect the compounds to move through the medium at different speeds only if the column is properly prepared. The need for auxiliary hypotheses such as these, changes the HD picture considerably. In the second step of the method, the deduction of observable consequences is only possible from the main hypothesis in conjunction with various **auxiliary hypotheses**. The falsity of an observable consequence therefore, via modus tollens, only implies the falsity of that conjunction. And a conjunction is false if any of the conjuncts are false. For example, if we do not observe three fractions condensing at different temperatures, all we can conclude is that the conjunction of the main hypothesis *and* the auxiliary hypotheses are false. The **conjunction** is false either if the main hypothesis is false or if the distillation apparatus does not work properly. Which of the conjuncts is false cannot be concluded only on the basis of the observations we make. More generally, we can say that we never test a single hypothesis alone, but only in conjunction with various auxiliary hypotheses. This is known as the **Duhem-Quine** thesis.

Auxiliary hypothesis: A hypothesis used to test another hypothesis, but which one does not intend to test, for instance background assumptions necessary to infer the empirical conclusion.

Conjunction: Two propositions joined by the logical operator AND. The conjunction is true if and only if both propositions are true

Duhem-Quine thesis: No hypothesis can be tested without the use of auxiliary hypotheses.

The Duhem-Quine thesis is named after the French physicist Pierre Duhem and the American philosopher Willard Van Orman Quine. The thesis considers not only assumptions about measurement instruments, but also, for example, basic assumptions such as that there exists a world external to the human mind, or that our sense organs do not systematically mislead us. We might be quite confident in some of these assumptions, but without them, we could not derive any observable implications from hypotheses. Occasionally, even such basic assumptions are shown to be false – as, for example, visual illusion tests teach us. The Duhem-Quine thesis implies that in order to falsify a hypothesis, we need to be *confident* that it is that hypothesis, and not its associated auxiliary hypotheses, that are responsible for an observation showing that the observable consequence is false. This leads us back to making judgments about confidence, however. Deductive inferences will clearly not deliver on these judgments. Consequently, the asymmetry between falsification and confirmation begins to break down. This, I think, is an important argument against Popper's suggestion for circumventing Hume's problem.

Falsification, although an important type of inference in hypothesis testing, faces the same problem of making sense of judgments of confidence and confirmation as inductive practices do. The previous arguments do not reject falsificationism altogether however. They only point out its limitations. Falsification might still be a viable inference method among others that scientists can use.

If scientists use falsification as an inference method, they must be particularly aware of the pitfalls of **ad hoc** modification. Consider the historical example of phlogiston theory. This theory, proposed in the 17th century, postulated that a fire-like element called phlogiston is contained within combustible bodies and released during combustion. One observable consequence of this would be that metals would lose weight if they burned, and thus release phlogiston. In conflict with this, it was observed that some metals gained mass when they burned, seemingly falsifying the phlogiston theory. But defenders of the theory replied that phlogiston might have negative weight, thus saving the theory from falsification. This reply from the phlogiston defenders is an example of ad hoc modification, as it does not serve any other purpose but to prevent the falsification of the theory. Of course, not all modifications are ad hoc. Many modifications of theories that were previously falsified often generated important new scientific insights. The criterion that Popper himself proposed is that a modification of a theory is ad hoc *if it reduces that theory's falsifiability*. It is those kinds of modification that must be avoided.

Ad hoc: The modification of a claim is *ad hoc* if (i) the claim has previously been falsified, (ii) the modification saves the claim from this falsification and (iii) it makes the claim less falsifiable – i.e. it does not allow deriving any new testable consequences.

Thus, to conclude, although Popper was right that, in the absence of ad hoc modifications, falsification is an important inference type in science, he overstated the importance of falsification in at least two ways. First, as seen from the Duhem-Quine thesis, scientists cannot avoid making confidence judgments about hypotheses so the strict asymmetry claim between falsification and confirmation is not correct. Secondly, despite its appeal in avoiding Hume's problem, falsification seen as a purely deductive program, is too austere a framework to fit the needs of scientific inferences made in practice. Instead, we need to explicitly address inductive inferences that aim to confirm hypotheses. This I do in the next section.

2.5 Confirmation

From the facts that H implies C and C is true, nothing can be deduced about the truth of H. This is not a deductively valid form of reasoning. It does not correspond to neither modus ponens nor modus tollens. Rather, any inference allowing us to say something about H, must *amplify* the information contained in the premises. Any inductive inference rule is fallible. You are likely to have made inductions that failed, be it when predicting the weather, when self-diagnosing an illness, or investing in the stock market. The interesting thing, though, is that different inductive inferences seem to have different degrees of fallibility. Some make us quite confident that the hypothesis is true, others less so.

Imagine, for example, you have a job testing products for a large sports retailer. You are currently checking a large batch of baseballs, and your hypothesis half have a cork center and the other half have a rubber core. Before you start looking at the batch, you hear from one of your assistants that he saw workers at one of the factories sewing baseballs both with cork and rubber centers, but couldn’t make out the proportion of each. If

you have no other information about this issue, a plausible guess is that the batch contains 50% cork and 50% rubber cores. But then you launch a large-scale test: you select 5,000 of these balls at random, open them, observe their core, and find half of them to be made of cork and half made of rubber. Now you still believe in the 50/50 hypothesis, but with more confidence. You still cannot be certain, of course: it isn't impossible that the balls in the batch that you didn't check have a different proportion of cork and rubber. But your degree of belief in the truth of the hypothesis, your *confidence*, should have gone up. Confirmation thus comes in degrees. This degree depends on several factors, for instance on the kind and quality of the evidence, as well as on the inference rule used.

Degrees of confirmation might be characterized *qualitatively*. The most recent assessment report of the International Panel on Climate Change (IPCC), for example, offers five qualifiers for expressing confidence in hypotheses ranging from "very low" to "very high". Alternatively, people have attempted to *quantify* confidence. The most prominent route here is to use probabilities to say that an observation O confirms hypothesis H. This means that the probability of that the hypothesis is true, given the observation, is higher than the probability that the hypothesis is false given the observation. For example, the probability that it has rained, given our observation that the street is wet, is higher than the probability that it has not rained, given our observation that the street is wet. This is, however, beset with a number of problems. First, not everybody agrees that it makes sense to assign probabilities to hypotheses because they interpret the concept of probability differently. One prominent position, **frequentism**, defines probabilities as the frequencies of repeatable observable events. Because hypotheses are not events, neither observable nor repeatable, frequentists argue that one cannot meaningfully assign probabilities to them. They are the wrong kind of thing to be said to have probabilities, the argument goes. Another problem is that probabilities are already used to express a property different from confidence. I might say, for example, that the probability of drawing a rubber core ball from the above batch is one half. You might then ask how confident I am in *that* claim, "The probability of drawing a rubber core ball is one half", and that is a separate question from the first one. Furthermore, in the above example, the probability of drawing a rubber core ball stayed the same, but the degree of confidence that this was the correct probability changed. Of course, one might answer the second question with a probability also, and have two distinct types of probability. But then one must keep these two numbers separate, which people often fail to do.

Frequentism: Probabilities are frequencies of repeatable observable events.

So specifying what we mean by degrees of confidence isn't trivial. But even if we agreed on a framework, it is still unclear what justifies confirming inductive inferences. To understand this more clearly, let's go back to the HD method. There, observing C to be true increases our degree of confidence that H is true. But why? What makes it so that observing C increases our belief in the truth of H?

The first answer is because H is *compatible* with C. When checking whether C was true or not, we could have made observations that

contradicted H, thus leading to it is rejection. But we did not. Therefore, we are now more confident in H. This compatibility argument offers only very weak justification of confirmation. First, one can deduce indefinitely many implications from a hypothesis, even though the truth of many of them seems highly irrelevant. For example, one can always deduce **tautologies** from anything: from the hypothesis "it has rained today", we can infer that "it has rained or it has not rained today" – and tautologies are always true. But surely, deducing a tautology from any H does not lead to an increase in our confidence in H. For another example, recall that from the hypothesis "Mammary cancer in mice is caused by X" one can deduce the observable consequence "Mice have mammary glands". But observing such mammary glands in mice does not confirm the hypothesis about cancer.

Tautology: A proposition or an inference which is necessarily true.

An additional concern with the compatibility argument is that very many hypotheses are compatible with any given observation. This is called the problem of **under-determination**. Take, for example, the problem of curve fitting. In curve fitting, we connect a finite number of observation in state space through a continuous line that is supposed to represent not only those observational points, but also the infinitely many points in between. We can think of the curve as a hypothesis that inductively generalizes beyond the observations. Now, does the fit of the curve to the observational points confirm this hypothesis? Well, maybe, but we need to admit that there might be another curve, distinct from the first, that is also compatible with the observations. And so are indefinitely many other ones. Each of them are substantially different hypotheses compatible with the observations but making different claims about the world otherwise.

Under-determination: An inference is underdetermined if there are multiple conclusion that would be equally supported by the premises.

A possible solution to this problem is to select the *simplest* hypothesis that is compatible with the observations. For the curve fitting case, this might work, as one might be able to rank these different curve hypotheses according to their degree of polynomial. But in many other under-determination cases simplicity ranking is not available. Consider, for example, the hypothesis that I have stomach cancer. Theory indicates that one relevant consequence of stomach cancer is heartburn. So, when I experience heartburn, should I substantially increase my confidence that I have stomach cancer? The answer is no. Very many hypotheses are compatible with heartburn. Thus, heartburn alone is not a strong sign of stomach cancer. A proposed solution for this problem is to introduce some criterion of *relevance* for observable consequences. Only the truth of relevant implications in the stomach cancer case, such as various effects of the cancer, would confirm the hypothesis. Specifying such relevant conditions is tricky, however, and requires a lot of background knowledge.

These examples shows that the compatibility of hypothesis and observation is not a sufficient basis for confirmation. We need something stronger. One proposal is that C confirms H because C would have been *very unlikely if H had been false*. That is, the possibility to observe C would have been remote if H were false and some other hypothesis true. This dependence of C on H however does not rule out that C is compatible

with other hypotheses. Rather, it acts as an additional requirement on whether a consequence can be confirming evidence of a hypothesis: only if there is such a relationship between C and H, does the observation of C confirm H.

The philosopher Deborah Mayo calls hypothesis tests that satisfies this condition **severe tests**. Severe testing solves our heartburn problem. Although stomach cancer is compatible with heartburn, the chance of suffering from heartburn if one has stomach cancer is not higher than the chance of suffering from heartburn if one has reflux disease or some other gastrointestinal disorder. Testing the cancer hypothesis by observing heartburn is thus not a severe test and observing heartburn does not substantially confirm the cancer hypothesis. Severe testing thus considerably improves the quality of inductive inferences.

Severe test: A hypothesis test is a severe test if the probability to observe a consequence would be low if the hypothesis were false.

Severe testing, however, requires a lot of information, in particular about alternative hypotheses and the probability of making the observation that one has made if those alternatives where true or false, respectively. Furthermore, this proposal does not solve all problems associated with confirmation either. What, for example, if one first made an observation and then constructed one's hypothesis in such a way that it would imply this observation? Such a hypothesis might meet the severe test condition, but would, arguably, not confirm the hypothesis. Consider curve fitting again. It seems illegitimate to construct the curve in such a way that it fits all the available information, then claim the curve as one's hypothesis and then use *the same data points* as confirming evidence for this hypothesis. Instead, only observations that have not already been built into the construction of the curve should be used in confirmation.

Finally, consider this case where severity seems insufficient to justify strong confirmation. The police used new breathalyzers, which falsely display drunkenness in 5% of the cases in which the driver is sober. However, the breathalyzers never fail to detect a truly drunk person. 0.1% of the population is driving drunk. Suppose a police officer stops a driver at random and forces the driver to take a breathalyzer test. The test indicates that the driver is drunk. How *confident* should the officer be that the driver is *actually* drunk? Try finding the solution yourself before continuing!

The correct answer is the probability that the stopped driver is actually drunk is about 2%. This might seem surprising. After all, the breathalyzers falsely display drunkenness in only 5% of the cases in which the driver is sober – that is, C would have been pretty unlikely if H had been false, and the breathalyzer still counts as a fairly severe test.

Yet, because the base rate of drunken driving is so low in the population, the relatively small error of a wrong indication becomes very influential. For example, consider a sample of 1,000 drivers. According to the information given above, only

Further reading

You can read more about Mayo's account of statistical inferences in her 2018 book:

Mayo, Deborah G. (2018). Statistical inference as severe testing: how to get beyond the statistics

0.1% - one out of these 1,000 - is driving drunk. And the breathalyzer correctly identifies him or her. But the breathalyzer also falsely identifies 50 out of these 1,000 as drunk although they are sober. Altogether, the test identifies 51 people as drunk, of which only one is actually drunk. One out of 51 equals roughly 2%. Thus, the probability of someone identified by a breathalyzer to be actually drunk is about 2%.

This is a case that meets the condition of a severe test, but it does *not* confirm the hypothesis *strongly*. It does confirm it a little, but just not as much as you probably thought. Mistaken judgments of confirmation based on these kinds of scenarios are called **base-rate fallacies**. Because the initial confidence in the hypothesis was so low, even such a severe test could not increase the confidence very much. In order to deal with such cases, we need a Bayesian approach to hypothesis testing. I will discuss such approaches later in the statistics section.

To conclude, confirmation of a hypothesis by an observation comes in degrees. To simply note that the hypothesis is compatible with the observation is not enough however. Rather, we need also to ensure that the hypothesis made the observation very likely, and that the observation would not be very likely if the hypothesis had been false. Such a severe test condition improves the quality of inductive inferences substantially, but even it faces a number of problems.

Part 3 - Observation and measurement

3.1 Different kinds of observations

Empiricism: sensory experience is considered to be the ultimate basis for knowledge.

Experience or – synonymously – observation plays a crucial and fundamental role in science. You can say, broadly at least, that all scientists are **empiricists**. Empiricists in the sense that derives from the Greek *emperia*: that sensory experience is considered to be the ultimate basis for knowledge. Today this might seem like the only serious contender for such a foundation. But that would be a mistake. Think of religious knowledge that many people, today and in the past, would claim. You find a beautiful example of that in the gospel of John. The disciples tell Thomas about Christ's resurrection, but Thomas says that he doesn't believe that Christ actually has risen and demand to see and feel the evidence for resurrection himself. So, Jesus meets with him and allow him to see and touch his wounds. Then he tells him off for needing such a lower kind of justification. According to the gospel, he says, "Because you have seen me, you have believed; blessed are those who have not seen and yet have believed".

So here is some non-empiricist contrast to this idea that beliefs are ultimately justified with sensory experience. I would say that all scientists will agree in their rejection of at least the second half of Christ's claim. In science, you are not blessed if you believe without having seen. However, what is interesting is that the contrast that we find in John is too simple in order to understand science. It doesn't mean that because scientists don't believe what they haven't seen, the only available alternative is to be like a doubting Thomas and require to see or touch for yourself in all cases. Instead observation, experience, is often much more complicated and indirect, and that's what I wanted to talk about now.

Direct observation: by unaided sense experience

Let us distinguish a number of kinds of observations. The first kind is the one that we find Thomas is after. He requires a **direct observation**: by unaided sense experience. He wants to use his own senses without any aids. Scientists often do that too: they see the colour of a liquid or they smell whatever particular gas exudes from the liquid, they might taste the acidity in the liquid (although you shouldn't try this at home), they may touch the vessel and see that it is warm, or they might hear it bubbling and fizzling. Those would be forms of direct observation and clearly scientists often engage with that.

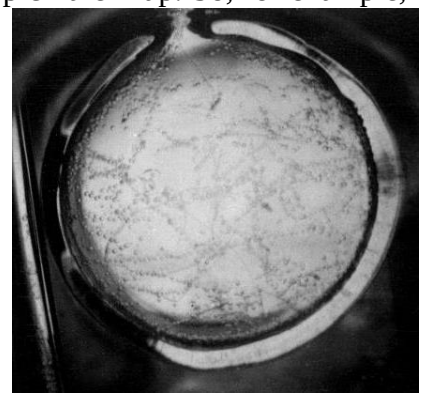
Aided observation: to amplify sense experience.

However, scientists do much more than this. For example, they use instruments, and they use them in different ways. Sometimes they use instruments to **aid** their direct observations. For example, they might want to experience bacteria or very far away stars, and those they cannot see with the naked eye, so they use instruments in order to amplify their senses. So, what is happening here? You cannot actually see protozoa with your naked eye but that light that is reflected by those little creatures is amplified or concentrated through a microscope in such a way that they become visible to your eye. Here, the important thing is that there is a *direct connection* from these organisms to your eye through an instrument that facilitates you actually taking up on that information. It is different

from direct observation, because it requires these aids, but there is still a direct connection between your senses and these objects that you are observing.

Indirect observation: sense experience of the effect of an event, but not the event itself.

Contrast that with another type of observation, which are called **indirect observation**. Many things that scientists are interested in are not directly observable. No matter how much you were to amplify whatever information is there, whatever signal is there, whatever causal base is there, it is not enough for our senses to pick them up. So, for example, we cannot directly observe subatomic particles. There is no optical microscope strong enough to see those. Instead, what we can do is that we can craft a particular experimental setup in which these objects that we are interested in have some effect that is observable. So, we are not observing these objects themselves, rather we are observing their effect.



This is the case for example in the bubble chamber, which was an early instrument to detect charged subatomic particles. You would have a superheated liquid, under substantial pressure, and you would then send charged particles through that. These charged particles leave an ionization trail. What they in fact do is leave bubbles in that superheated liquid. It is those bubbles, or those trails that the bubbles create, that are observable. That is what you see in this photograph on the right-hand side. You see these little traces here. Now, scientists say that they are observing particles. And I agree that they are observing them, but it is important to note that this is a different kind of observation than either using your senses, with or without any aid. Here you are actually seeing the effect of these subatomic particles, not the particles themselves. And that is why it is called indirect observation: it is mediated through this intermediate step that we are then able to detect with our senses. And of course, this chain might be longer, and we might in principle distinguish between degrees of indirectness, but we do not need to do this here.

To summarize, I have given you different kinds of sensory experience, and I have argued that the distinction here is based on the different way of how the object of our interest is related to our senses. And I have also argued that instruments play a quite different role, either as aids in direct observation or in this setup of detecting effects of the objects that we are interested in.

Theory dependence: An observation is theory dependent if it requires theoretical assumptions in order to make and justify observations.

3.2 The theory dependence of observation

In the previous section, I distinguished three kinds of observation, and now I will use this distinction in order to argue that at least some observations are **theory dependent**, that is they require theoretical assumption in order to make and justify these observations. Why is theory dependence important? Well, it is important because it complicates a very prominent but probably false view about empiricism

or a version of empiricism, called **logical empiricism.** It was an early 20th century philosophy of science that had substantial influence on the sciences, in particular modern physics. Their idea was that we can separate out observations from theory so that we can have a separate realm of observational sentences, created from observations we have made. These were seen as a separate body of knowledge, from which we can then build and justify our theories. And that is of course very convenient, because you then have a clear foundation on which all theories are built, and when in doubt whether a theory is justified, you just have to look up in the set of observation sentences whether there is sufficient evidence for that theory to be justified.

Logical empiricism: Theories are created by inferences from fundamental observational statements.

That sounds very nice, but theory dependence makes things more complicated. I will now argue for that this theory dependence is a problem for logical empiricism. First of all, let us look at indirect observation, for example in the bubble chamber experiment. If we want to say that we actually observe subatomic particles in these bubble chamber trails, and be justified in saying this, then we need to justify the assumption that there is a close connection between these trails and the particles that cause the trail. We have to make assumptions about the properties of the liquid in the superheated state, about the relation between particles and the ionization, and also about the relation between the particles and the size of the bubbles. Without making these assumptions, it is entirely unjustified to point bubble chamber trails and say, "I see particles". These assumptions form a theory, and the truth of this theory establishes the required relation between the property or the event of interest and the observable property, and also its regularity, saying that charged particles sent through a bubble chamber always, or at least most of the time with sufficient probability, produce these trails and nothing else does. Indirect observations are always *theory dependent* in this way. We are not justified in claiming that we have observed something indirectly without also assuming that the underlying theory is true.

This is one example of the theory dependence of observations, but now I want to argue that even aided observations sometimes require theory. For this, let me give you the historical example of Galilei Galileo observing the four moons of Jupiter. This is considered to be one of the great moments of history of science. It is often described as Galileo's observations through the newly developed telescope, as inescapable evidence for the four moons; and the reaction of his critics simply as the stubborn scientific and intellectual establishment refusing to even consider Galileo's evidence. This view of the event is described in the German playwright Berthold Brecht's play "Life of Galileo", where a cardinal of the inquisition says to Galileo: "I could look through your telescope, but I don't need to". The implication is that if the cardinal only had looked then he would have been convinced and there would have been no more question that Galileo was right.

However, here is a report of someone who was there when Galileo made his first rounds among the scientists, and who experienced how difficult how difficult it actually was to determine the quality of Galileo's

empirical evidence. This is a letter from Martin Horky to the astronomer Johannes Kepler: “Galileo Galilei, the mathematician of Padua, came to us in Bologna and he brought with him a spyglass through which he sees four fictitious planets. ... On Earth it works miracles; in the heavens it deceives, for other fixed stars appear double. … I have as a witness most excellent men and most noble doctors … all acknowledged that the instrument deceived. And Galileo became silent, and on the twenty-sixth, a Monday, dejected, he took his leave from Mr. Magini very early in the morning.” (Letter from Martin Horky to Kepler, April 1610). The report of this eyewitness of this new instrument for aided direct observation tells us that the common narrative is false. It did not only require to look through the telescope with an open mind to be convinced. Intellectuals and scientists at the time who did look were often *not* convinced.

How was this possible? The telescope exhibits optical aberrations, an optical phenomenon caused by the properties of the lens where light is spread out rather than focused on a point. When looking at objects on earth with the telescope, you can go and directly unaidedly observe the object. You walk away a certain distance and say “aha, it actually shows the object, however, there are certain blurring’s that occurs on the rim of the vision field – aberrations”. However, when you point your telescope to a very distant star you cannot go to that star and observe it directly. Now, you see that things are blurred and things appear double and that there seems to be all kinds of effects that the telescope produces. So you question if it actually depicts the stars in the heavens in the right way, and you cannot really check.

As we see, there it is uncertainty for Galileo’s contemporaries whether this instrument produces apparent experiences that do not correlate to any real object or if it is accurate, and therefore they rejected it, at least in this initial encounter. So, what do you need in order to counter this argument? You need a theory of optics. You need to explain why optical aberrations occur and how you can distinguish these aberrations from the depiction of the actual phenomena that amplified by the telescope. Given that there was no established theory of optics at the time, the scientific community was at least to some degree justified in being cautious of this observation, that so strongly contradicted established knowledge at the time.

Clearly, this does not apply to all aided observation, because in many cases you will be able to compare what you see through the instrument with actual looking at things directly. This is why they said “on earth it works”; they found the telescope useful, but not for the purpose of looking at stars. To determine if the measurement equipment actually shows reality or deceives you need theory, which Galileo didn’t have. So at least in some cases, even for aided direct observation, a theory is required. And that makes both indirect and aided direct observation theory-dependent.

If some types of observations at least sometimes need theory, then there is something wrong with the picture that all theoretical sentences are derived from observational sentences. We cannot have observations as

the foundation of theory, because now observation already relies on theory. Instead of having one thing built on the other, you have some kind of a circle. Circularity need not always be bad news for justification – but surely it poses a problem for those who seek justification though a secure foundation.

This is actually one of the most interesting methodological and philosophical problems: how to get around the problem of making observations to support theory if the observations already require theory. Here is the example of temperature measurement. Developing reliable measure instruments for measuring temperature was a huge thing in the 18th and 19th century. Observing temperature will, of course, be an indirect observation, at least in the sense that we want to have an exact numerical value on the temperature. One typical way that was proposed in the 18th century and that has dominated for quite a while is to construct a measurement instrument for temperature by linking temperature to density of liquids. These thermometers are typically constructed by using two-point measures, typically boiling and freezing point of water, and by assigning numerical values to these two points. These assignments are quite arbitrary – for example in the old thermometer that Celsius used, seen to the right, the numbers are inverted: hundred for freezing and zero for boiling[1]. You then divide the scale in equidistant steps and you state that as the liquid expands, it rises in the glass column and the temperature equals the different numbers. This requires that there is a linear relationship between temperature change and density change. How do you know this? Well, this is a theoretical assumption. What evidence would you have for such an assumption? Experiments involving temperature measurements! But that is exactly what you are trying to figure out, so the reasoning seems circular.

More generally, such cases have been described as the *problem of nomic measurement*: we want to measure quantity X. Quantity X is not directly observable, so we infer it from another quantity Y, which is directly observable. But in order to do so we need to rely on some regularity that relates X and Y, that X can be expressed as a function of Y. The form of this function cannot be discovered or tested empirically because that would involve knowing the value of both Y and X, and X is the unknown variable that we are trying to measure. That is a more general formulation of the circularity problem.

How does one deal with such a problem? Certainly not by trying to find a secure foundation! That the problem is circular has already been established, so there is no arguing with that. A more promising route is to translate the problem into a decision problem, asking what alternatives you are choosing between. In other words, if you want to indirectly measure temperature (and thus generate sufficiently precise quantitative

---

Foto: Peter Häll, Tekniska Museet

measurement results), what alternatives do you have and what is the best alternative to choose? If you were a 19th century physicist, you would have already hit upon the qualitative observations that various liquids and gasses expand with rising temperature, so you take this as a promising route. But then you have many more choices to make; which liquid or gas should you fill your thermometer with – should it be water, alcohol, mercury, air, or something else altogether?

It was this choice that the French scientist Henri Victor Renault (1810-1878) focussed on. He tried to show that certain kinds of thermometers generated more consistent measurement than others. By minute variations in their respective setup he showed that thermometers filled with some liquids varied enormously and those filled with others varied a lot less. For example, he found that mercury thermometers that were set up with different kinds of glass showed a wider variety of readings than air thermometers made with the same range of glass. Similarly, mercury thermometers under different pressures in the vials had a wider spread than air thermometers under similar conditions. He argued that because the air thermometer showed lesser variation, it gave more consistent, more comparable results and therefore was the better choice for measuring temperature.

Of course, Renault did not solve the problem of nomic measurement this way – at least not in the sense that he somehow justified the linearity assumption between temperature and liquid/gas expansion, without invoking this very principle in temperature measurement. But he showed how one could systematically go about selecting the best of such instruments, if such a linearity assumption held. That work helped a lot in making temperature measurement more precise, which lead to improved data, which led to the development of more refined theory, which facilitated the development of new measurement techniques. And because each of these new techniques gave results that were consistent with the previous ones, we now are highly confident in the assumptions involved in each of these measurement processes – thus finally resolving the problem of nomic measurement, at least for the case of temperature. Renault thus did not address the nomic measurement problem head on, by trying to find a foundation. Rather, he started working to improve one part, conditional on some unproven hypotheses, and then others added their work, which fit with Renault's findings, so that a larger and larger coherent mosaic evolved. This is an example of a *coherentist approach*, which I discussed in the context of Hume's problem in chapter 2.

To summarize, I have argued that indirect observation is always theory dependent, while aided direct observation is sometimes theory dependent. Theory dependence complicates empiricism, you cannot go with the foundationalist picture, but because of this we are facing the circularity problem and I have given you and example of how this can be solved.

3.3 Operationalization

Even for those events or properties that in principle can be observed directly, like acidity or temperature, in science we often prefer an indirect

observation of these properties. We do not measure the comparative speed of cars with our eyes standing at the goal post but rather we use a timer and a distance measure. We use thermometers of different kinds for the measurement of temperature and we use pH indicators for acidity. For these cases, we are now relying on a more involved process for which we have, or believe we have, justification in terms of justified theories or models. Why do we do that even in those cases even in those cases we *could* rely on direct observations?

There are two reasons, first that direct observations are often unreliable. There are plenty of cases where we can show that our direct observation gives the wrong result. For example, length or height perception vary with observational distance, and loudness perception varies with background noise. Therefore, it is not very reliable to have people use their ears to make comparative judgements on loudness, instead it is more useful that you use an instrument that records the effect of an audio signal and displays this in an indirect form.

Furthermore, as we will discuss later, direct observation typically only give qualitative properties. When I argued that temperature and acidity in principle can be directly observed, it was also clear that we could at best get qualitative ranking, but we wouldn't be able to say much about distances between different observations, because we cannot quantify. Instead, we need an indirect observational process for that. That we get when we operationalize. To **operationalize** a property of interest is to provide a way of linking it to a directly observable effect. And this has two parts: we have the **property of interest** and we have some effect that we can directly observe. We are then linking these through a **hypothesised causal chain**, or sometimes a correlational relationship. This chain or this relationship must be sufficiently reliable. If that is the case, we can infer that because we are observing the effect and because we have hypothesised that relationship, we conclude that the property of interest was present or absent to a certain degree. For example, when measuring temperature, we are hypothesising a linear relationship between liquids' density and temperature, and between density and height in a column, and we are then directly observing the height of the liquid in the column. With attitudes, beliefs and desires in humans we are assuming a causal relationship between the presence or absence of certain of these attitudes in people and the answer to these questionnaires, and then we take these answers as giving indication of these attitudes.

This can also go wrong. You can operationalize in a way that is not good, that actually fails in important ways. Here is familiar example: using a spring scale to measure the mass of an object. Note that on this spring scale on the left side, it says kilograms. This is even imprinted on this device; it claims that it indicates mass. However, what is the relationship between mass and behaviour of the spring scale that is required here? What is needed is a direct proportionality between the mass of the object and the gravitational force exerted on the object – and that assumption is false. That spring scale goes terribly wrong if we use it on the moon. Nor will that spring scale work if you use it on an accelerating train. Still, for

Operationalization: To operationalize a property of interest is to provide a way of linking it to a directly observable effect through a correlational relationship or a hypothesised causal chain.

Property of interest: The property that is to be measured.

Hypothesized causal chain: The supposed link between the property of interest and what can be directly observed.

the marketplace or for you weighing your luggage before you are going to the airport, it works fine. It works fine for certain areas, but you must be careful not to use it in the circumstances where this relationship does not hold anymore.

Or take the measurement of people's voting decisions. When you ask people about their intended voting behaviour, you are assuming that there is a regular relationship between their (directly observable) answer to the pollster and their true voting decisions. But that assumption is false: depending on perceived social norms, fads, trends, etc. people will sometimes try to conceal their true decisions and lie to the pollster. Which arguably is one of the reasons pollsters were so wrong about the outcomes of e.g. the 2016 US presidential election or the 2017 Brexit referendum. A lot of people were apparently not willing to express or reveal their intended genuine voting behaviour, thus leading to wrong forecasts. They said one thing but voted in a different way. The operationalization thus failed.

Because operationalizations often fail, we need to specify a number of quality criteria for operationalizations. In the first place, we need to properly *define* what the property of interest is. What do we mean by happiness of an individual human, for instance: an emotional state, a judgment, a number of physical properties? If we disagree on the definition of what we want measure, then of course we should not be surprised if the results on our measurements are a mess.

We must also have a valid relationship between the property and the effect. It must be the case, according to our best knowledge, that the directly observable property shows up *only* when the target property is present, and it *always* shows up when the target property is present. Furthermore we also must be sure that there are not too many disturbing factors involved, for instance in the case with the spring scale, there are many circumstances were we would say that it does not work because there are just too many factors that disturb the proportionality between the mass and the behaviour of the spring scale. That is what I mean with sufficient stability for the purposes at hand. We must have stability of this relationship under changing background conditions.

Finally, we should pick a property that actually is directly observable with *sufficient precision*. We don't want to operationalize and spend a lot of time in crafting such an operationalization, when we end up with a property that is still difficult to observe, or that other people cannot observe, or that cannot be observe with sufficient precision. All of these are important for making a good operationalization.

Let me end this section with a caution regarding a related but different idea, which is the program of **operationalism**. Operationalism is the particular view that all basic properties of interest are *defined* through the operations by which they are measured. So in contrast to what I said above, the definition of the target properties are not given independently, but are given by the way we are measuring them. Take a psychological

Operationalism: The view that all basic concepts are defined through their operations, by the way which we aim to measure them.

example: what is happiness? Operationalism answers that happiness is whatever the standard measures of happiness measure. So if we measure it by recording various responses to certain stimuli, then happiness *is* the behavioural responses a person shows when exposed to certain positive stimuli. For another example, consider length. Operationalists define length by the operations through which it is measured, so e.g. as that property that is measured with a ruler. So you ask "what do you mean by length" and you point to the ruler, or the standard meter, and you answer, "Whatever operations you perform with that ruler, the results you get – that is what you mean by length".

In the first place, note that one can operationalize all kinds of target properties one is interested in measuring *without* agreeing to operationalism. As soon as you propose a definition of the target property that does not refer to the measurement process itself, you are on your way to crafting a non-operationalist operationalisation. So operationalism is a particular view about how operationalisation should be crafted, and a controversial one, too.

The main motivation for an operationalist view is that it avoids the theory dependence of measurement. If you declare the property of interest to be that which your measurement operation measures, then you don't need to argue that there is a stable relationship between target property and directly observable effect. That is already build into your operationalist view of what measurement is. While that is an undeniable advantage of operationalism (theory-dependence, as I argued, poses important problems for good measurement, after all), I think the disadvantages of such a position are sufficiently many to reject this view.

The first disadvantage is that operationalists multiply concepts that shouldn't be multiplied. The operationalists suggest that if we measure length or temperature in different ways, these are different concepts. So, temperature measured with a mercury thermometer is not the same concept as temperature measured with resistance thermometer, because they are measured differently and thus defined differently. Or length, measured on the nanoscale, with particular operations is not the same as length between galaxies, measured in very different operationalizations. So instead of length as one concept, we now need to differentiate $\text{length}_{\text{nano}}$, $\text{length}_{\text{galaxy}}$, $\text{length}_{\text{mid-size}}$, and so on. This does not seem correct, rather I think it is very intuitive to say, "We have one concept of length or distance, and it doesn't break into different concepts with different measurement methods".

The second disadvantage is that by embracing operationalism, we lose an important ability to critically evaluate operationalisations. If we agree on a common (non-operationalist) definition of happiness, then there might be multiple ways how to operationalize the measurement of this property; and because they all aim to measure the same property, we can compare these variants with one another. The operationalists, in contrast, can always say, "Well, if you measure it this way, then that is just a different concept of happiness". Then, we have as many happiness concepts as we

have measurement processes, and no way to compare them: my survey questions are different from your survey questions, or your way of evaluating facial expressions is measuring a different concept than my way of evaluating the concept of happiness. That is a recipe for avoiding critical communication, which is so central for the enterprise of science. Instead, one should admit that we are aiming at the same property, happiness; we are just doing it in different ways – some of which might be better for a given purpose, and some worse. Furthermore, we might learn that independent ways of measuring the same concepts yield the same results. Such *measurement robustness* results are important evidence for the quality of our operationalization, as the case of temperature measurement shows.

The practices of temperature measurement shows that operationalism does not a dominant view in the sciences. Scientists define temperature independently of any measurement processes as mean kinetic energy. They do not say, "Temperature is what we measure with mercury thermometers". Instead, they give theoretical reasons for why there is a stable relationship between temperature and the expansion of certain liquids, electric potential, electrical resistance, thermal radiation, etc., and each of these relationships then justifies operationalizing temperature (the property of interest) in terms of some directly observable effect identified by these relationships. That the results produced by these different operationalizations are approximately the same then has given further support for these theoretical reasons, confirming the quality of these operationalizations.

To summarize, indirect observation requires operationalizations, and scientists even in those cases where we have possibilities of directly observing often prefer to operationalize these properties none the less, and I have given you a number of quality criteria for operationalizations, and argued in particular that operationalism is not a way to improve operationalism.

3.4 Measurement

What is special about measurement? Measurement expresses the result of an observation quantitatively – i.e. represented by numbers. A measurement process consists of four main steps, two of which we have already discussed. First, we define the concept to be measured, and then we specify an operation that is linked to directly observable property, and we argue why this link is valid and sufficiently stable. However, for something to be a measurement, we are doing two additional things: specifying a measure, and by defining the measure's units, determining what type of scale this measurement is using. A measure is an object of comparison, for example a tape measure, a balance scale with weights or a mercury column. The unit of a measure is the quantification of the measure property to be compared. Examples of such units are the meter (m), the gram (g) or degree Kelvin (K).

Comparability: A measure must be comparable to what is being observed.

Now what is needed for a good measure? In the first place, a measure must be **comparable** to what is being observed. Start with the simple case,

distance measurement for mid-sized objects. In such distance measurement, we are performing a measurement on a direct observation. We are observing an object directly, and then we are observing how many times our measure fits next to it or besides it. In order to function as a measure, this thing must have the same property as the object we are observing – i.e. it must have an extension, so that we can lay it next to the object of interest. That's why we use rods, sticks, tape, etc. to measure length, but not gasses, emotions or numbers – the latter do not have a (clearly defined) extension.

Things are a bit more complicated when we measure properties that are operationalized and observed indirectly. In temperature measurement, we don't require our measures to have mean kinetic energy. That is because we compare the measure to the directly observable effect of the operationalisation of temperature – e.g. to the height of the mercury column. So the measure must have an extension, because that is how temperature was operationalized. Because there are many different operationalizations of temperature, there are many different measures: besides extension, others measures involve electric potential, electrical resistance, or thermal radiation (even though these often are themselves operationalized).

The other quality criterion for a measure is its **stability** over time. A tape measure that changes its extension with air humidity, for example, would be really useless. This has actually been a problem for most of the comparative units, like the meter or the kilogram, which used to be actual physical prototypes that exemplify exactly, or as precisely as possible, that measure. Now, if you use that kilogram regularly in comparing the mass of other objects, then it will lose some material – it will actually lose mass. It has been a problem in everyday practical applications in historical times that weights actually lost some of its mass, and therefore its measurement changed, and this was the case for the prototype as well. We need that the measure is sufficiently stable.

Stability: A unit remains unchanged over time.

Probably because of these many requirements, what we see in the history of measurement there is an enormous effort to try to find the right measure. There have been many attempts to find the right kind of distance measure; it started out from being defined as certain distance between two points on the earth's surface, it then went on to the crafting of prototypes, then it finally moved on to natural constants. If we choose a natural constant, we do not need prototypes at all – instead we are using a theory to identify certain underlying magnitudes, that we find repeated in large number of phenomena. It is complicated to do this, but when this was successful, the uncertainty associated with these different measures decreased by orders of magnitudes. And that is one of the most crucial goals, that we don't have instability, variations in the units every time we use it. Decreasing that variation has been the purpose for which an enormous amount of effort has been put into it. Very recently, November 2018, the international system of units was changed when the last prototype, namely the prototype of the kilogram, was replaced by a natural

constant. Until then the kilogram was defined with referenced to the object that sits somewhere in a vault in Paris.

Once one has found the right measure, one needs to specify the measures **unit**. The unit quantifies the measure's property, and thus allows counting it. By declaring "this rod is a meter", and then observing that one must lay 20 end-to-end to connect to points X and Y, one can conclude "the distance between X and Y is 20 meters". How big or small these units are is entirely conventional. There is nothing natural about the meter being the meter, or the kilogram being the kilogram; it is merely set by agreement. Of course not everybody agrees, that is why we have meters vs. feet, Celsius vs. Fahrenheit – different units that quantify the same properties of interest, with the same measurement processes.

However, while some aspects of a unit are set conventionally, others are not. For example, some measures like the Meter and the Gram have a natural zero, others, like degrees Celsius, do not. For some measures, one can replace one quantification quite freely with another, while for other measures, such a replacement, if possible at all, is more restricted. The reason for this lies in the differences of the properties that the measures are intended to represent. This is typically expressed by categorizing different measures into different scales.

The most common scales are the ordinal, the interval and the ratio scale. In an **ordinal scale** all you do is you order objects according to a qualitative comparison with regard to a particular property. Take for examples Mohs scale of mineral hardness, ranging from 1 to 10. If a sample X is given 8 on this scale, you are justified in concluding that it is harder than a sample Y that received 4. But you are not justified to conclude that X is "twice as hard" as Y, or that a sample Z with hardness 6 has the same difference in hardness to X as to Y. Nor does it make any sense to say that assigning 0 to a sample means that it has "no hardness" in any absolute sense. Due to the nature of the property hardness, and due to the way how Mohs scale operationalizes it, none of these additional conclusions can be justified.

Ordinal scale: Order objects according to a qualitative comparison with regard to a particular property

That is different for example in an **interval scale**, which orders objects, but also quantifies the distance between them. The Celsius and the Fahrenheit measures are such interval scales. If a sample X measures 8C and a sample Y 4C, you are justified in concluding that X is hotter than Y. You are also justified to conclude that a sample Z with 6C has the same difference in temperature to X as to Y. However, one cannot conclude that an object with 0C "has no temperature"; and for the same reason, one cannot conclude that X is "twice as warm as" Y.

Interval scale: orders objects, but also quantifies the distance between them

Such conclusions, however, are justified for a **ratio scale**, like temperature measured in Kelvin (K). Kelvin measures the amount of mean kinetic energy in an object. Thus, to say that an object has 0K implies that it has no kinetic energy whatsoever – and an object with 8K is indeed twice as hot (i.e. has twice the kinetic energy – as an object with 4K.

Ratio scale: orders objects, quantifies the distance between them but also has zero point where zero represents an absence of the property.

Which scale you are measuring a property with has important implications for how you can interpret and how you can process your measurements. This is often ignored in scientific practice. Take for example the Likert scale, a psychometric scale commonly used in questionnaires, which asks subjects to respond to various statements by choosing a reply from “I strongly disagree” to “I strongly agree” with a number of steps in between. Researchers often assign numbers to these response options, and then use such a quantification to calculate the mean response from a sample. But calculating the mean requires interpreting the scale as having meaningful distances – e.g. “Agree” is three times as distant from “I strongly disagree” as “I tend to disagree” – so the Likert is assumed to be an interval scale. Yet the respondents might have thought of it simply as a ranking, without any sense of the intensity of their agreement, thus treating it as an ordinal scale. In such a case, conclusions derived from the mean of a Likert scale would be hard to justify.

One way how to clearly distinguish these scale types is by the transformations they permit. A scale transformation is the replacement of the numerical representation of one set of observation with another numerical representation, *without* changing any of the empirical content of this representation. For example, we could decide that Mohs scale should have a range not from 1-10, but from 10 to 1000 (let us call it Shmohs), and that furthermore, the first 9 levels should be represented in increments of 10, while the last (because diamonds are so special) should get the number 1000.

Such a transformation:

Mohs		Shmohs
1	->	10
2	->	20
…		
9	->	90
10	->	1000

would yield a Shmohs scale that conserves the ranking of any samples measured in Mohs: if X is harder than Y, because according to Mohs(X)>Mohs(Y), then if you follow the above transformation rule: X is harder than Y, because according to Shmohs(X)>Shmohs(Y). This holds for any transformation rule that preserves the ordering property. Such transformations are called *positive monotone.*

All positive monotone transformations are permitted for ordinal scales, but not for interval or cardinal scales. If you apply the above transformation rule to a set of temperature measurements in degrees

Celsius, you *do* change the empirical content of this representation. That doesn't mean that Celsius measurements cannot be transformed at all – for example one can transform any Celsius measurement into a Fahrenheit measurement without changing its empirical content. But such transformations are more constrained than positive monotone ones. Instead, only those transformations are permitted that multiply the scale with a factor >0 and that add any positive or negative number to it. Such transformations are called *positive linear transformations*. The above transformation is not positive linear, but the transformation from Fahrenheit to Celsius is:

$$T_C = 5/9\ T_F - 32.$$

Ratio scales, finally, allow only an even more restricted transformation: they can only be multiplied by a positive factor, but cannot have any absolute term added or subtracted. These are called *positive scalar transformations*.

Considering these scale choices, it becomes clear that the quantification of any observation is a highly constructed process. It involves making decisions what to measure, and how, and with what means. And it also requires to justify these decisions. The philosopher Patrick Suppes aptly expressed this insight in this way: "We cannot literally take a number in our hands and apply it to a physical object. What we can do is to show that the structure of a set of phenomena under certain empirical operations the same as the structure of some set of numbers under arithmetical operations and relations." The "structure of the phenomena" is the property we want to measure. The "structure of some set of numbers under arithmetical operations and relations" is the scale type that we choose for this. These two aspects must match with one another, and we are not justified to perform mathematical operations on the measurements that go beyond this match.

3.5 Measurement error

However well constructed, measurement can go wrong. But we can specify measurement error and can take steps to reduce it. Measurement error is broadly defined as the difference between a measured value and the true value of that quantity. This is a philosopher's notion of error, because after all, we do not know the true value. We are not able to simply say, "We have the true value, now we have the measurement, now we can say that there is an error". We can only conceptualise that this is what we mean by measurement error, but we then have to think about ways how we can possibly find out about it.

Systematic error: errors caused by specific factors that confound the measurement process. It is also called bias or inaccuracy.

We can distinguish between two broad kinds of error: systematic and random error. **Systematic error** is caused by specific factors that confound the measurement process. It is also called bias or inaccuracy. For example, a scale might be miscalibrated, because the calibrating person inadvertently used the wrong standard weights when calibrating the scale - he mistook the 200mg test weight for a 100mg test weight (let's assume he was a lazy calibrator who only checked the correctness of zero

and of one weighing – which he then messed up). That creates a (huge!) systematic error in the scale: if it is otherwise accurate, this error will show up in every weight measurement performed with it, and it always will be of the same magnitude. Someone using this miscalibrated scale, unless they have independently obtained information about the mass of the objects they are weighing, will not have any reason to suspect that something is wrong. The scale will give consistent measurements, in the sense that it correctly ranks objects according to their mass, it gives the same mass for objects being measured repeatedly, and it will give different mass measurements for objects of different mass. Once the systematic error has been detected, however, then one can correct the measurement process and remove the source of this error – in this case, by correctly calibrating the scale. One can even correct the error in already performed measurements, by subtracting the effect of the miscalibration from the readings, thus getting the correct result.

The scale example illustrates a number of important features of systematic error. First, it is caused by a specifiable factor – in this case the miscalibration. Had this factor been properly controlled (i.e. had the calibration been performed correctly) the systematic error would not have occurred. Second, as long as the confounding factor is active, the all measurements with such a process will always include the error, and they will include the erroneous effect to the same extent. The miscalibration does not produce varying deviations from the true value – no, it is always the same direction (it will measure all objects as having less mass than they actually have) and it will always do so to the same degree – 50%.

Whether the confounding factor is active will in many instances depend on the background conditions. For example, an electronic measurement instrument might only exhibit a systematic error when a certain degree of humidity causes a short in one of the circuits. Then this instrument will only include this systematic error when operated in sufficiently humid conditions. But holding such background factors stable, systematic errors occur regularly. Measurements that do not suffer from systematic error are called **accurate**.

Random error, in contrast, consists in the fluctuations of measurements, due to factors that cannot or that for practical reasons will not be controlled. A typical example of random error is noise in an electronic signal. The causes of such noise vary greatly and typically are not known for particular signals. However, one can associate particular kinds of noise with certain kinds of devices, thus reducing and at least changing the type of nose by switching to a different device. Besides this lack of knowledge about the specific cause, the most important characteristic of random error is that it has a zero expected value – if you take multiple measurements, each measurement will be around the true value. They will all individually deviate but taken together, as you are approaching an infinite amount of measurements, you will actually get the true value – as long as no systematic error is present. Measurements that do not suffer from random error are called **precise**.

Accuracy of measurement: Measurements that do not suffer from systematic error.

Random error: Errors caused by the fluctuations of measurements due to factors that cannot or that for practical reasons will not be controlled.

Precision: Measurements that do not suffer from random error.

We can make these two kinds of errors, and the relationship between them, clearer with the following simile. Consider a shooting target, consisting of concentric rings. Its centre, the bullseye, represents the true value of a property to be measured. Now let our "hits" on this target be the individual values we get from applying our measurement process to the same object under the same circumstances (i.e. doing repeat measurements).

If we have precise measurements that suffer from a systematic error, then the "hits" will all be moved away in the same direction and to the same extent from the true value. We get a bundle of hits to the south west of the true value, because something is pushing these measurement values away from the true values. In this case, the dart illustrates an inaccurate but precise measurement. Measurement procedures that does not suffer from systematic errors are called accurate or valid.

If we have accurate measurements that suffers from a random error, the "hits" will be located around the true value. They will all individually deviate but taken together as you are approaching an infinite amount of measurements, you will actually get the true value – if your measurement is entirely accurate but imprecise. A measurement that doesn't suffer from such random error is called precise or reliable.

How do we deal with these different errors? Random error is relatively easy to deal with given that we know that it is a pure random error, since it then has an expected value of zero given sufficiently many measurements. Because we can average over all these different points, that measurement error will disappear. That is why you should always repeat your measurements. How many repetitions are required depend on what your measurements are about, how strong the random error is. This does improve precision, but it does not improve accuracy.

How do we deal with systematic error? The first step is to identify that the measurement suffers from systematic error, and to identify the factor that causes it. If we can identify it, if we can predict what the systematic error is, then we can eliminate the cause or counteract it when registering the result, and we will improve accuracy.

Identification of systematic error is not trivial - if your measurement only suffers from systematic error, you will have a relatively tight bundle of repeat measurements and you might assume that they are all at the true value. So simple repetition of measurement is not going to help you here. One way to check whether that assumption is true or not is to vary background conditions. Perhaps by varying temperature or humidity, you trigger a factor that causes systematic error, or you switch it off. Then you get different results from repeat measurements under varying conditions, indicating that your measurement suffers from some systematic error.

Another powerful strategy to search for systematic error is by comparing different ways of measuring (i.e. different measurement processes) applied to the same target property. Think of temperature: we have an

independent definition of temperature as mean kinetic energy. We have many different ways how to operationalize this property, many ways to connect mean kinetic energy to something directly observable. We can now look and see whether we get the same measurements from different operationalizations. If we measure the same object, under the same conditions, with different, causally independent, instruments – mercury thermometer, Thermistor, Thermocouples infrared thermometer – and always get (approximately) the same results, then the best explanation of this convergence might be that none of these instruments suffers from systematic error. This inference is called **convergent validity**. Getting the same readings from operationalization of the same concept that are independent from each other is one way to support your conclusion that you have correctly measured the property.

Convergent validity: Convergent validity is achieved if several, differently operationalized in causally independent ways, measurement processes, all applied to the same target under the same conditions, give the same result.

However, it might be a bit premature to be satisfied with convergent validity alone. The best explanation for why we get the same results from different measurement processes might be that they all are free from systematic error. But here is another one: they might all be equally blunt – i.e. non-sensitive to the target property. This is different from the claim that they all have the same systematic bias, which we excluded by requiring that the different measurement processes must be causally independent. For example, I can use a thermometer and a scale to measure the acidity in a liquid, and they all will give the same result for the same liquid under identical circumstances. But it would obviously be wrong to conclude that therefore they did not suffer from systematic error. Instead, the "convergent" result here is only an artefact of their non-sensitivity to acidity. So in addition to convergent validity, we also need to check whether a measurement instrument, when applied to different targets, yields different results. This is called **divergent validity**.

Divergent validity: Divergent validity is achieved if an operationalization yields different results when measuring substantially different targets.

Only when we have shown that a set of causally independent measurement processes exhibit divergent validity individually and convergent validity collectively can we conclude that they are free of systematic error. If either of these conditions is not satisfied, you know that there is some systematic error somewhere, and you will need to search for where it is – e.g. by varying various background conditions that you suspect might have an influence. Where to look, and what to look for, will very much depend on the specific domain in which you are working, the kind of theory available to you, and the kind of experience you have with the measurement instruments. There is little I can say about this in general.

In conclusion, let me point out the connection with what I have said today, in particular what I said about measurement error, to the previous section on hypothesis testing. In a previous lecture, I discussed the complications of hypothesis testing by the involvement of auxiliary hypothesis. Auxiliary hypotheses make it difficult to identify whether falsifying evidence actually concerns the hypothesis or the auxiliary hypotheses. And what I have shown you today are all examples of auxiliary hypothesis. All assumptions about the observation, about the measurement instrument, about the potential ways that something can go wrong comes into the hypothesis testing structure are auxiliary

hypotheses. They will all be important for determining what the empirical implications of a hypothesis are and when testing that empirical implication. When the implication is found to be false, that is, the empirical consequence does not appear, it will go back to both claims about the observation and measurement process, and the hypothesis. That is why it is crucial to identify what errors there might be in measurement and how likely they are.

Part 4 - Experiments

4.1 The Fuse-box Experiment

Imagine that you are making toast one morning. When pressing down the lever on the toaster, the entire kitchen goes dark – a fuse blew. You know that an electrical fuse is blown when the electrical system is overloaded, perhaps by a short circuit. You can simply replace the fuse, but if the cause of the blown fuse remains, it will just happen once more. You therefore need to experiment. That is, you need to observe the circuitry, but not just passively observe. Rather, you need to actively manipulate the circuit, in order to both intervene on the target variable and in order to control background variables. These activities of **observation**, **manipulation**, **intervention**, and **control** characterize the process of experimenting, and distinguish experiments from other observational practices.

Observation: Registering an empirical phenomenon.

Manipulation: Changing a factor in a study.

Intervention: Changing an **independent variable**.

Control: Reducing the influence of disturbing factors from influencing the outcome.

Independent variables: Variables that are hypothesised to affect the **dependent variable**.

Dependent variables: The variable that changes depending on the changes to the independent variable.

To determine what happened, you start by checking which appliances are connected to that fuse and are turned on – let us say the toaster, a washing machine and a lamp. They were all turned on yesterday, but the fuse wasn't blown then. Today however it was, and all appliances stopped working, although they were still switched on. This observation alone does not help you very much. In order to understand what just happened and in order to repair the circuit, you will need to identify what in particular caused the fuse to blow. There are many potential causes for this.

H1: Toaster blew fuse

H2: Washer blew fuse

H3: Lamp blew fuse

H4: Toaster & washer jointly blew fuse

H5: Toaster & lamp jointly blew fuse

H6: Washer & lamp jointly blew fuse

H7: All three appliances jointly blew fuse

H8: Other factors blew fuse

One hypothesis type is that any one of the individual appliances broke and caused a short circuit—this is expressed in hypotheses number one to three. Another type is that operating the appliances jointly caused and overload—as expressed in hypotheses number four to seven. Last, they are, of course, many other potential causes of the fuse blowing, but because we cannot check them individually in this particular system, they are collected in the hypothesis number eight.

In order to test for these particular causes, you first need to establish control. That is, you switch off all particular causes identified in your hypotheses one to seven: all three appliances. You then replace the fuse so that the circuit has current. If some other cause that caused the fuse to blow

still was active - for example, of faulty wiring, then we should at this point, observe another blow. This observation would support hypothesis number eight. (Note however that even if we do not observe another blow in this setting, we still cannot rule out hypothesis eight – some other cause (e.g. a current spike) could have caused the earlier blow, but is not operating anymore).

Now, to test each particular hypothesis, you intervene on the target variable identified in the hypothesis and switch it on. For example, you test hypothesis 1 by switching on the toaster. You rule it out if you observe that switching on this potential cause does not make a difference in the circuit. If you do not observe any effect, you can rule out the hypothesis and the same with hypothesis 2. You then test hypothesis 3 by switching on the lamp while, of course, having switched off the other potential causes. And now, you observe that the fuse again blows. And based on that, you can accept hypothesis 3. You have identified the lamp as the cause.

Now, you might think you were done – the lamp was the cause of the blown fuse. However, accepting hypothesis 3 does not exclude that other causes are also operating. Several things might have been broken at once. From here, you will go on in the same fashion, testing the other hypotheses. So now, you would intervene on pairs of appliances, switching them on at the same time, while controlling for background variables by keeping the third appliance switched off. In the end, if you have properly thought of all possible hypotheses to the broken fuse, your experiment should justify you to accept at least one of them.

To summarize, **experiments** like the above are a type of observational process characterized by three features. First, the control of background variables, achieved through manipulation of the experimental system. Secondly, the intervention on target variables identified in the hypotheses you are testing through manipulating the experimental system. And thirdly, the observation of differences produced by intervention. These features distinguish experiments from other types of observational processes.

Experiment: An observational process characterized by control of the background factors (often through manipulation), intervention on the real target variable through manipulation and observation of the difference caused by this intervention.

But what makes this type of observation process so special for science? I mentioned before, that experimental observation justifies accepting or rejecting certain hypotheses. The English philosopher, John Stuart Mill, provided an interesting analysis of this kind of reasoning, which he called, the **method of difference**. Let us have a look and see whether it helps us understand why experiments provide such a justification.

Method of difference: By creating two identical groups, and making an intervention on only one of them, the causal effect of that intervention can be studied, since the intervention is the only factor different between the groups.

The reasoning starts by stating our question (1): we have observed a phenomenon E and now want to know what caused it – for example, what caused the fuse to blow. Then, (2) we make an informed guess that some factor C is the cause of E, which forms the hypothesis to be tested. In simple cases like the fuse box case above, your guess will be informed by an intuitive understanding of the electrical circuit. In more complex cases, you will want to consult theory. Next, (3) we start the experiment by

1. We ask: what causes phenomenon E?

2. We conjecture: C causes E (hypothesis)

3. We produce two situations S1 and S2, in which neither C nor E occur and in which all causally relevant factors are the same. (successful control)

4. We activate C in S1 but not in S2. (successful intervention)

5. We observe that E occurs in S1 but not in S2. (observation of differences)

6. In S1, something causes E (from 5)

7. In S2, nothing causes E (from 3)

8. The only differences between S1 and S2 are C and E (from 3 and 4)

Conclusion: C causes E

producing two identical situations S1 and S2, in which neither the factor C nor the factor E occurs, and at which all potential causes are controlled. This is sometimes difficult to do: in the above case, there is only one circuit to experiment on, so you cannot generate *two* identical situations. What you need here is description of a type of situation, which can have multiple instantiations. Think of two identical fields for a plant growing experiment, two identical batches of concrete objects for a durability test, or two groups of patients with identically distributed properties for a drug trial.

Then (4) we intervene on C in situation one [S1] but not in situation two [S2]: for example, fertilize one field but not the other, coat one batch of concrete objects but not the other, or inject one patient group with a drug, but not the other. By intervening in S1 but not S2, we produce a difference in causal factors between these two situations. We then (5) observe whether this intervention produces any difference in outcome between S1 and S2, and more specifically whether this outcome in S1 amounts to the phenomenon E. Thus (6) in S1, something must have caused E, while (7) in S2, there was no observed change and therefore no cause should have operated.

The reason for having this **control group** S2 is that there are many factors that can influence plant growth, concrete weathering or patient health. If in those cases one only had *one* group, one would not know whether it was the intervention or some other factor that produced this effect. But with a control group, if we have constructed our experiment properly, the only causal difference between these two situations is the intervention on C in S1 (the **treatment group**), and the absence of that intervention in S2 (the control group), and the only difference in outcome is the presence of E in S1, and the absence of E in S2. Thus we conclude that (8) in S1, E was caused by C, and in S2, the absence of E was caused by the absence of C. Hence, we accept the hypothesis, C causes E.

In the fuse box experiment, however, we did not have a control group. Rather, we had to (implicitly) assume that there were no other factors that could have caused a change in the system at the same time as we performed our intervention. For a household electric circuit, this might be a reasonable assumption (and furthermore an assumption we had to accept – there simply is no way for most homeowners to create an identical

Control group: A contrast class that is not subjected to the intervention, but is similar to the treatment group in all other relevant aspects.

Treatment group: A class that is subjected to the intervention.

control circuit). But for most systems, such an assumption is highly dubious, and would lead to a substantial reduction in experimental quality.

Mill's method of difference reconstructs the reasoning behind testing hypotheses with experiments. It shows how experiments can provide strong evidence for accepting or rejecting hypotheses. But it also shows where this reasoning might go wrong - because in our reasoning, we make assumptions that might actually be false. For example, we might fail to control all relevant factors. If that's the case, the factor C will not be the only active cause in S1, and the reasoning is no longer valid. We might also fail in our intervention on C and not make the difference that we claim we would make. Or we might fail to observe correctly, claiming a difference in outcome where, in fact, there is none. Finally, we are making some deeper assumptions about the nature of the causal relations in the system we are experimenting with, and these assumptions might also be false. Clearly, arguments from experiments can be fallible in this way. They might contain error because our assumptions about the experiment might contain errors. Mill's method helps us to detect these potential sources of error. Knowing these sources will help us to design better experiments.

4.2 What are experiments?

For a scientific practice to constitute an experiment, it needs to consist of control of the background factors, intervention on the target variable by manipulation, and finally observation of the result of this intervention. However, not all observations in science are experiments. In fact, there are whole areas where scientists do not and cannot experiment. In astrophysics, for example, you cannot experiment because you are unable to manipulate the relevant objects that you are investigating. You cannot manipulate stars. So instead, you just have to watch what the stars are doing anyway. Astrophysics is therefore not an experimental science. And in many other scientific areas, we cannot experiment either, sometimes for the sheer physical impossibility, sometimes just because it is too expensive, and in many cases because it is unethical or illegal. You cannot just manipulate the world in which people live and have rights – rights to property, to physical integrity, to respect and dignity. And sometimes, you just have to resort—quite often, actually—back to just observing what's going on. In those cases, we are talking about **observational studies**. Observational studies are not experiments.

Observational study: An observational process characterized by control of the background factors and observation of an outcome, lacking an intervention.

Here is an example from the social sciences which I think is interesting, which starts out from the intuitive idea that dangerous jobs offer higher wage, compared to less dangerous jobs. The hypothesis thus says that "H: Wage increases with professional health risks". The investigation would then go and look at different workers and the different wages and look at their different risk of death. The problem with this, of course, is that there are lots of other factors that influence wages and that these vary with different workers. You will most likely not find workers who differ in wage and then *only* differ in their risk of death. They also differ in education, age, gender, type of job, etc. And all of these might at least

potentially be influences on their wages. Yet the hypothesis claims that health risks affects wage when other things are equal. So we need to somehow isolate the influence of risk on wage from all these other varying factors. How can that be done? In an experiment, such factors are background variables which you control it by manipulating the system. That cannot be done here – it would be unethical and illegal to change workers' wage and risk exposure, and impossible to change their age, gender, education, etc.

In this observational study, you cannot control the background variables by manipulation, and you have to deal with them in some other way. One way is by employing statistical methods, which make relatively strong assumptions about whether the influence of these respective variables can be isolated by analyzing larger data sets. That is very different from being able to experimentally manipulate the background conditions in such a way that they are the same, or that you are even able to eliminate the influence of some of these factors all together.

An observational study will most often only allow for a lower degree of control – but it is important to note that an observational study is not devoid of control all together. When observing a star with a telescope, for instance, we can make sure that the telescope does not misrepresent what we see, and that we make an observation on a clear night. When determining the risk of death in a profession, we can make sure that the data on salary is correct by collecting and comparing self-reported salary levels as well as official numbers, and similarly with death rate.

There is a special kind of observational study where you might be able to argue that there is the same level of control as an in experiment. That is often, and quite confusingly, called a **natural experiment**. But look out – compared with our definition of an experiment, this is not actually an experiment. The example that I have here is from early epidemiology. There was a cholera epidemic in 19th-century London. One doctor, John Snow, developed a new method of investigation: for the worst affected neighborhood, he observed where the infected lived. For each observation, he made a mark on a map. He then looked at these two groups, the infected group and the non-infected group. And he argued that these two groups were basically identical in terms of the distribution of their potentially relevant properties—their age, their gender, their professions, whether they were sick otherwise, what their income was, and so on. Snow argued that there was no difference in potential causal factors except for one, which could be identified through his mapping: where they fetched their water.

Natural experiment: An observational study where circumstances are organized as if there had been manipulation for intervention and control but no manipulation is in fact done.

At that time, typically, people did not have indoor plumbing. But instead, people would go to a well nearest to where they lived to fetch water for cooking, cleaning and drinking. Snow discovered that some of these wells drew their water from a spot on the Thames downstream from London – so that the unfiltered sewage of London was part of what people living close to these well were consuming. And it was these wells that the infected had regularly consumed from. Other wells, however, drew their water from elsewhere, without being contaminated in this way – and most people

consuming this water had not been infected. Hence, Snow concluded, the difference in infection must be explained by the *only* factor by which the infected and non-infected differed: it was the polluted water that caused the Cholera infections.

Today, we know about the germ theory of disease and therefore we know that diseases are transmitted through water supply amongst others. That theory did not exist in Snow's time. But his use of Mill's method of difference, applied to his observational data, allowed him to arrive at this causal conclusion that then led to changes in water and waste water management, and thus to substantial improvements in public health.

Note, however, that the good doctor did not perform any *manipulation*. He just went about and observed. He was just lucky that he could not find any other difference in any of the background variables except for this one—water supply. So therefore, he did not have the same problem as the one that I showed you here where all these other factors varied in all kinds of other ways. Against this stable and identical background, he could argue that the only difference was the water supply. And that then had to be the difference maker, the cause of this effect, i.e. the disease. So we can say that natural experiments are a kind of observational study where we cannot perform manipulation and where we cannot intervene on the target variable but where we find that all relevant background variables are successfully controlled, even though this was not achieved through manipulation.

Now, contrast that with, for example, a **field experiment**. In a field experiment, we are often taking the background variables as given. The only thing we are manipulating is the intervention. An obvious example is an experiment testing the impact of fertilizer on plant growth. When setting up treatment and control groups, one identifies two agricultural areas with identical properties. But in a field experiment, one only selects such areas with identical properties; one does not manipulate them. For example, one doesn't change the soil quality, or the sun exposure, or the watering conditions – one just takes whatever these conditions are in one's geographical location and makes sure that they are the same for both treatment and control group (so that we do not test the fertilizer e.g. only on that part of the field that gets a lot of sun while the other part is in the shade, for instance). So in a field experiments, most of the background conditions are taken as given and not manipulated. Most of the manipulation happens only in the intervention (e.g. in fertilizing the treatment area but not the control area). Note though, that a field experiment does not have to be performed outdoors or in a field – the term signifies the level of control in the experiment.

Field experiment: An experiment where there is manipulation for intervention, but where several background variables are not controlled, in particular not controlled through manipulation.

In a **laboratory experiment**, to the contrary, we are manipulating all, or practically all, of the background conditions. We are massively manipulating the system, perhaps by more or less building things up from the bottom. We consider all variables and determine for example at what temperature, at what pressure, at what pH value, etcetera, we are performing our intervention. The difference between the lab and the field

Laboratory experiment: An experiment where there is manipulation for intervention and where all or most relevant variables are controlled.

would then be how much we are engaged in manipulating the background conditions. Again, as with field experiment, a laboratory experiment – despite its name – does not need to be performed in a laboratory, although a laboratory often offers conditions to control the background factors efficiently.

Another thing that is often called an experiment but is not technically an experiment, are simulations: a **simulation experiment**. This often means programming a computer to perform a certain sequence of steps, which is used to represent something. For example, we might *represent* a landscape, certain wind movement on it, and then we might represent some kind of pollution that is emitted *in a computer model*. We then observe how the pollution distributes in that simulated landscape. We have control of the background factors – extreme control, even, since there are no background factors we have not programmed (i.e. manipulated). We also have the intervention we programmed and can observe a result – so why is not this really an experiment? The important thing here is that the control and the intervention is done on a representation of the real world, rather than the real world itself. We are manipulating, but we are manipulating only the representation in the computer: the causal factor determining the observation we get is not the wind, but an equation, describing the wind. We are not manipulating the target itself. And that distinguishes simulations from field or laboratory experiments.

Simulation experiment: A model study where the model is created as to mimic an experiment, but where the variable that is intervened on is only represented, rather than actually present.

I have given you a couple of examples, some of which are experiments and some of which are not. Let me now propose a definition of experiments that helps distinguish these cases – identifying field and laboratory as experiments, while identifying natural and simulation experiments as non-experiments. “An experiment is a controlled observation in which the observer manipulates the real variables that are believed to influence the outcome, both for the purpose of intervention and for the purpose of control." A real variable is a factor that is actually out there, and is for instance not a representation of a variable

4.3 The Purpose of Experimentation

Experiments often provide more powerful evidence than other observational processes, but experiments can also fail. I will now discuss different kinds of errors in the design of experiments that destroy or at least hamper their evidential power. Let me start with correcting a widespread misunderstanding about what a “failed experiment” is. There is a group of so-called failed experiments that are not really failures at all, because the experimental design does not contain any grave errors. Perhaps the best example of these is the famous Michelson-Morley experiments performed in 1887. Its goal was measuring the difference between the speed of light traveling in different directions. Michelson and Morley expected such a difference because they, like most other scientists at the time, believed that the vacuum in space was filled with a substance that allow light waves to pass through it. This was called the Lumiferous Ether Theory. They designed a very clever and well-controlled experiment to test this. Nevertheless, it is often considered the most famous failed

experiment because Michelson and Morley, against their expectations, could not detect any relevant difference between the speeds. They thus failed to do what they *intended*. However, it was exactly because the quality of the experiment was so high that it stood up against their expectations and the dominant theoretical view. The Michelson-Morley experiment, rather than being a failure, turned out to be very productive by providing the first strong evidence against ether theory.

Many so-called failed experiments are actually productive in this way. Learning when an expected effect does *not* occur is often as important as learning when it does occur. However, not all experimental failures are productive in this sense. There are experimental observations from which we cannot conclude anything of interest. In particular, if the experimental process is not designed correctly – for example, if the conditions of Mill's Method of Difference are not satisfied – then it does not produce evidence that justifies either acceptance or rejection of a hypothesis. It is these kinds of failures or errors that I want to discus in this section. The most important errors in the experimental design concern the lack of control of the relevant causal factors. These errors can arise in many different ways.

Let us have a look at some prominent examples. The Flemish alchemist Van Helmont investigated what made plants grow. For this purpose, he planted a small willow sapling in a vat with soil of a specific weight. He let the tree grow, giving in only water, and after five years weighed both soil and tree again. While the soil had not changed in weight by what he could measure, the tree had grown more than 30-fold. As Van Helmont had observed, the addition of water as the only difference, he concluded that the entire mass gain of the tree was produced by the water. Today, we of course, know that this conclusion is false. Much of the mass of a tree comes from the air—more specifically, the carbon dioxide it processes via photosynthesis. Helmont could have found out about this if he had *controlled* the supply of carbon dioxide. But he did not have the technology to do so, and more importantly, he *did not even consider* it as a relevant causal factor. At his time, most scientists took air for a static substance that could not turn into solids like wood. He drew the wrong conclusions about his hypothesis because he failed to control for an important background factor. That is a very common experimental error, as we often do not know what the causally relevant factors are even today.

Experimental errors compromise the process of experimentation. Those processes fail to control for background variables, or they intervene not only on the target variable but also on some other factors, or they fail to correctly observe what is to be observed. Only if we manage to avoid such errors can the observations justify their conclusion. In other words, only then do we say that the conclusions from the experiments are **internally valid**. But how can we know whether our experiments include errors? We have two general strategies here. First, we should apply all available knowledge when designing experiments. This includes theoretical knowledge—for example, about the composition of air, about heat flowing from bodies, including measurement instruments, and so on.

Internal validity: An inference about an experimental system is internally valid, if the relation between intervention and observed effect inferred from the experiments is indeed true and is not confounded by uncontrolled background factors.

More than that, there are two main strategies scientists employ to reduce the number of errors in experiments.

The first strategy to reduce experimental error is to be mindful of the experimenter's unwanted kinds of influence in the experimental process. One kind of error arises when experimenters fail to control factors that arise from or with observation. The so-called **observer effect**, or **influence problem**, describes how the act of observing itself might have an effect on the target variable. For example, an experimenter inserts a thermometer in a liquid to measure its temperature, but the heat contained in thermometer changes the temperature of the liquid. Or, to take a different example, when running a behavioral experiment with humans or animals, the experimental subjects may be aware that they are being observed, and therefore change their behavior. These additional factors must be controlled, or the experimental result will not justify any relevant conclusion.

Observer effect / influence problem: The act of observation changes what is observed.

The **confirmation bias**, or **interpretation problem**, describes psychological effects on the person observing the result of measurements. For example, if the person knows which samples have been intervened on and expects that the intervention produces a strong effect, then there often is a detectable tendency to read off larger effects from these samples than from those not intervened on. Unless one controls for this bias, the experimental observation will not justify any relevant conclusion.

Confirmation bias / interpretation problem: The observation is registered incorrectly due to psychological properties of the observer.

In experiments with humans or animals, problems also arise from additional effects from the intervention. **Placebo effects**, for example, arise when the patients' belief in the effectiveness of the treatment intervention makes a difference. They might show as a lessening of symptoms, affected through their belief that the treatment is effective, even if this treatment does not contain any causally active substance or procedure. To control for this, modern medical studies always assign some patients without their knowledge to a control group where they receive such inactive placebo pills.

Placebo effect: The therapeutic effect that an inert substance or treatment, designed to have no therapeutic value, has on a patient

When dividing samples into treatment and control group, **selection bias** is another kind of error that threatens. Inadvertently, one's selection might be influenced by the very factors that have an influence on the target variable. A famous example comes from an observational study performed in World War II (I am including this here, although it is not an experiment – can you tell why not? – because it so well illustrates this failure of control, that also arises in experiments). The Royal Air Force tried to minimize their bomber aircraft loss from enemy fire, and therefore investigated how to make the airplanes stronger. They surveyed the damage of all returning aircrafts and recommended adding armor to those areas that showed the most damage. However, mathematician Abraham Wald argued that what one should really consider were those aircrafts that did *not* return from the mission. Those were the ones that sustained fatal damage - the ones that did return showed damage that was not fatal. Instead, armor should be added to areas that were *not* damaged in

Selection bias: A factor influences the selection of the sample, or the division in to test and control group, which one did not aim to create.

returning aircrafts, assuming that damage is uniformly distributed across aircrafts.

Knowing about such errors are a part of the theoretical training of every scientist. However, a scientist also has a lot of tacit knowledge – knowledge of things that are learned from experience in working in a lab but that many people might have a hard time expressing explicitly. Which liquid to use for chemical extraction, at what temperature, are often more dependent on the experienced laboratorian's intuitions, transferred through generations perhaps, and not inferred from any specific theory. The same for knowing how best to clean your laboratory equipment. Experimenting is very much also a craft. But this means that avoiding experimental error is not merely an issue of finding the right theoretical knowledge. It also requires practicing a lot with people who have experimental experience.

The second strategy of detecting errors start is to recognize that science is a collective endeavor, and an important part is testing hypotheses that have already been tested before: **repetition**. Now not all experiments are repeatable. Only if enough information is available about the setting, the intervention, the control, and the instruments used can we say that an experiment is repeatable. We then might have different reasons to repeat an experiment. One goal is to **reproduce** the previously reported results. If a repetition yields the same result, then we can be more confident that it did not contain errors. If, however, the results cannot be reproduced, this indicates that the previous or your own experiments contains an error and you can start searching for it. Another reason for repeating an experiment, this time with slight variation, is to see how dependent the result was on the previous experiment setup. For example, you might repeat a chemical reaction between two liquids, but now, at a different temperature than in the original experiment. We call such processes **replication**. If you fail to replicate an experiment in this way, then you learn that certain factors like, for example temperature, have a causal impact on the experiment result. These are ways how to find out about experimental errors, in particular failures of controlling casually relevant factors. Once you learned about them, you can then proceed to correct the error by performing various strategies of control.

Repetition: Performing a study again, exactly following the description of the original study.

Reproduction: Repeating a study and obtaining the same result as in the original study.

Replication: Performing a variation of a study with the intention to achieve the same result, where some aspect of the study has been changed to learn how this aspect influenced the result.

4.4 Experimental Control

Experimental control consists of two parts. First, we need to identify the background factors relevant for a particular phenomenon – I discussed that in the previous section. Second, once we know what these factors are, we have to look at how we are performing the actual control, how we can influence the system so that we get the relevant background conditions in place. Here we have a number of different strategies.

The first strategy is to use a divide between control and treatment group. This means taking a sufficiently large sample on which we want to perform the experiment, and then dividing that sample into two or more groups. We then performing the intervention only on one group, while the other group we do not perform the intervention on – as it was described in Mill's Method of Difference. Now, this sounds pretty easy, but it rarely is.

In medicine, for example, we know that there are large placebo effects. If someone gets a pill, irrespective of whether that pill has an active ingredient in it or not, receiving the pill might already make them feel better. To make sure that the *only* difference is the active ingredient, the control group also gets a pill, just without the active ingredient in it. That is one way how to correctly divide control and treatment group.

Now we can see why it is important to use the control treatment group division instead of, for example comparing observations made before an intervention with observations made after an intervention (as for example in the fuse box experiment). Let us say we want to see whether changing the lighting in the rooms at KTH is going to affect student performance. After doing the intervention and changing all the lights, we compare last year's results with this year's results. You can probably see that there are plenty of other potential background factors that might have changed between last year and this year – for example a new student cohort, a changed curriculum, new teachers, etc. If instead we are changing the lights only in the control group (e.g. the L and M buildings), and comparing the change in performance for the courses taught here with the change in performance in courses taught in control group (e.g. the E and V buildings, where we did not change the lighting), then we can expect that those background factors that change from year to year (new student cohort, a changed curriculum, new teachers, etc.) are affecting *both* treatment and control group, and thus cannot make any difference between them. The difference in performance that we do observe can therefore with higher confidence be attributed to the intervention. While there is still more one could do to control other potential sources of error, the simple division into treatment and control thus improves the internal validity of the experiment beyond that of the before/after study.

Once we have divided the group into control and treatment group, we need to make sure that the background factors in those two groups are the same. In some cases, we might not even need to hold factors **constant** because we are able to **eliminate** some of the background factors. So, for example, when we are interested in the falling behavior of bodies, air resistance is one of these background factors. We can, of course, perform falling experiments by keeping air resistance constant. But it might be better if we removed air resistance altogether for example by constructing a tower in which one create a vacuum. Other examples include the elimination of radio waves in a Faraday cage, or the elimination of gravity by moving the experiment far enough into space.

A special case of elimination is **blinding**. I have already mentioned placebo effects: I know that I am getting a treatment or I believe that I am getting a treatment that might have an effect on how well I feel, and we therefore it is important that we ensure that the subjects involved in experiments do not know whether they are getting a treatment or not. Because of this, we are leaving the subjects in the dark about whether they get a treatment or not: that is what we call subject blinding or **single blinding**. Single blinding is thus one way to mitigate the observer effect or influence problem that our intervention creates.

Constancy: Holding background factors constant between test and control group or between trials in a study.

Elimination: Removing the influence of a background factor

Blinding: Eliminating observer effect or the influence problem through limiting information about the study to participants or observers.

Single blinding / subject blinding: Blinding the participants in a study, for instance about whether they are in the test or control group.

But then we also have the effect from the experimenter side. The experimenter might read the instruments differently when they know whether this is an instance from the treatment or control group. This need not be a case of intentional deception: experimenters might genuinely expect big differences to occur in the experiment, and due to their expectation, make and interpret their observations in slightly different ways. In order to prevent this confirmation bias or interpretation problem, it is important to also **blind the experimenter**, and even maybe the person who is analyzing the data, from knowing which individuals were in the treatment group and which ones were in the control group. This is called experimenter blinding.

Experimenter blinding: Blinding the observers in a study, for instance about which subjects are in the test and control group.

In experiments with human or animal subjects, there are thus two kinds: subject and experimenter blinding. Where both are implemented, the literature calls **double blinding**. Obviously, in many natural sciences, there are no experimental subjects and you cannot double blind. Therefore, one might simply use the term blinding, without saying single or double, meaning experimenter blinding only.

Double blinding: A study is double blind if both subject blinding and experimenter blinding are implemented.

A further form of controlling or performing control is **separation**. We cannot always eliminate the influence of a factor, but we can often to separate a factor from a number of different factors and register its influence independently. Take for example the Gravity Probe; an experiment that sent a hydrogen maser, a highly accurate clock, into space to measure time dilation with high precision – i.e. the rate at which time passes in a weaker gravitational field. This was compared to a control group of clocks on earth's surface. But in order to implement the intervention (the placement in a weaker gravitational field), the clocks had to be rapidly moved vertically away from earth. That inevitably created a Doppler shift in the treatment group, a factor affecting the frequency measurement, which could not be eliminated. But the experimenters could separately *measure* the Doppler Effect, and subsequently "calculate out" the Doppler Effect from the time measurements. That allowed experimenters to *separate* these two factors and only get as a result the time dilation.

Separation: Registering the effect of a background variable.

4.5 Randomization in Experiments

Randomization is an important feature of many experimental processes. There are good reasons for why one might want to randomize one's experiment, which I will explain, but there are also some inflated claims and expectations about what randomization can achieve, and I will try to debunk these.

But first, what is **randomization**? When dividing samples of experimental objects or subjects – e.g. chemical substance, bacteria cultures, material samples, patients, school classes – into control and treatment groups, experimenters face the question how to divide them. Randomization offers a systematic answer: you use a random process to assign them to either treatment or control. You can do this by flipping a coin or using a lottery machine or a computer. Furthermore, you can conceal the assignment so that neither the experimental subjects, if you are experimenting with

Randomization: Using a random process to choose a sample or to divide into test and control, for instance by flipping a coin.

animals or humans, nor the experimenter knows who is in the treatment and who is in the control group – that is, randomization makes blinding easier. Experiments that are set up this way are called randomized controlled trials (RCTs). Note that randomization in RCTs refers only to the division process, not the sampling process. That is, an experiment is randomized if the division of the sample relies on a random process, even though the way you sampled from the population might be non-random and even highly biased!

Randomized controlled trials have acquired a very high methodological status, particularly in medicine. Many scientific organizations consider RCTs or systematic reviews of RCTs to be the highest quality evidence available - because of randomization. Such ranking of evidence is called **evidential hierarchies**, and are distinguished by the processes through which they are produced. In extreme cases, this hierarchy has led researchers to focus only on RCT evidence and disregard lower ranked evidence if these types of evidence conflict.

Evidential hierarchies: A ranking of types of evidence based on the processes that produced this evidence.

Saying that one type of evidence is better than another requires justification. So, what are the arguments for taking randomization so seriously? There are at least three good reasons for why one might want to randomize one's experiment. The first is that it helps eliminate bias in the control/treatment assignments. If the experimenter performs the division herself, she might, intentionally or not, assign those particularly responsive to the treatment group and those less responsive to the control. This would threaten the internal validity of conclusions from the experiment. Randomization replaces the experimenter's influence and thus prevents such a bias. Second, even if there is no threat of such a bias, randomization might still be a good idea because it might help the experimenter to convince others that the treatment/control division was not rigged. Finally, randomization is a useful instrument to realize blinding – i.e. preventing both subjects and experimenters to know who gets a treatment and who does not. For these three reasons, it is typically a good idea to randomize experiments in particular because randomization typically does not cause any harm.

However, although randomization is helpful in these ways, there are other practices that can perform equally well. All that matters for eliminating assignment bias is that the assignment to treatment or control is not correlated with any relevant factor that must be controlled (if it is correlated, then the purpose of the division into treatment and control is lost). Whether such a correlation exists or not will depend on the particularities of the case. Take for example the durability test for concrete objects mentioned above. Imagine the producer of these objects delivers them in crate a 50 pieces. One can randomize their assignment: take out each object, flip a coin, and assign it to the treatment group if the coin falls heads, and to the control group when it falls tails (note that this procedure might lead to unequally sized treatment and controls, but that shouldn’t concern us here). Alternatively, one could argue that the order in which they are packed into the crates is not correlated to any factor relevant for their durability. If that is true, then the following assignment

procedure will also be unbiased: unpack each crate in a particular order (e.g. start from the upper left side), assign the first, third etc. object to treatment and every other one to control. That's of course just an example – there are hundreds of more rules that also might be unbiased. In any case, it is a good idea to make such a rule explicit when using it (just as one would make explicit if one randomized), so that others can confirm for themselves whether the assignment is indeed not correlated with any relevant factor that must be controlled. Making this explicit will also convince others that the experimenter had no undue influence. And, of course, there are many other blinding strategies. Thus, although a helpful instrument, randomization is not a necessary practice for any of these objectives.

It is sometimes claimed that randomization ensures that background factors are equally distributed in treatment and control groups. This is not true – unfortunately. Because if it were true one would not have to worry about controlling background factors anymore. One would just randomize, and the job were done. To understand why this doesn't work, consider a die. If we roll a die a *very* large number of times, the results will converge to a distribution in which each face comes up an equal number of times, and the average score count will be 3.5. However, for any smaller number of die rolls, this long-term distributional property does not hold. When, for example, rolling the die 10 times, we should not be too surprised with this result with sixes showing up more often than other faces and an average score being five. And while the convergence to 3.5 is pretty good after several hundred rolls for a normal six-sided die, one would need a much larger number for a 120-sided die. Most RCTs are like such many-sided dice with an insufficient number of rolls to reach convergence.

As an example, let us say we have a sample of bacterial cultures, with different pH values. We suspect that the pH value influences the effect of our experimental intervention, so we need to control for it. Assume we know that pH values in bacterial cultures are normally distributed around seven. So, the claim goes, if we randomly assign the bacterial cultures to treatment and control, then we will get identical normal distributions of pH value in the treatment and in the control as well. But is this really true? No, because any given randomization might distribute all high pH samples into the test group and all low pH samples into the control group. We should not be surprised that at least sometimes a randomization yields control and treatment groups in which the mean pH value differs substantially. In fact, any particular randomization might yield such differing distribution. The claim that background factors are equally distributed is true only for infinite sequences, but experimenters typically randomize only once and certainly not infinitely many times. The British statistician Ronald Fisher, who strongly argued in favor of RCTs, was nevertheless aware of this problem and warned that in practice experimenters will find the distribution to differ widely. We must therefore conclude that RCTs do not guarantee equal distributions of background factors in the control and the treatment groups.

Experimenters are, of course, aware of this and therefore check for such imbalances once a randomization has been performed. For example, if they find pH values to differ substantially between treatment and control after randomization, then they will randomize again. Alternatively, they are using their background knowledge about potentially disturbing factors and control for these factors before randomization. For example, they might divide the sample into sub samples with particular pH values and then randomly assign objects from each of these sub samples to either treatment or control group: **data stratification**. These strategies can only be applied to correct for imbalances of **known factors**. However, these imbalances will also arise for **unknown factors**. But because they are unknown, one cannot correct for those by stratification. Thus, randomization cannot guarantee the control of unknown factors.

Another problem arises from the confusion of random assignment and random sampling. A sufficiently large random sampling from a population is a good method to establish the representativeness of that sample. Experimental results from randomly selected samples therefore can be used to make good inductive inferences to properties of the population. But as I pointed out above, an experiment being an RCT does not imply that it is based on random sampling – only that its sample has been randomly assigned to control and treatment group. Of course, being an RCT does not necessarily exclude random sampling, it just isn't a required part of the package. So in practice, many RCTs do not involve random sampling. Consequently, the (valid) conclusions one obtains for the sample might not be valid for the whole population. In particular, a result that is valid for one sample might not be valid for a different sample drawn from the same population.

In medicine, for example, the sample of patients is typically constituted by those who at a given time walk into the hospital and exhibit the relevant kind of symptoms. It might be in principle possible to perform a sampling process that is representative, but random assignment (what is called randomization in the RCT) itself does not ensure that. Consequently, we cannot simply transfer our RCT results to other situations unless we argue that those situations are relevantly similar to the ones under which we perform the RCT. Such a similarity judgment, however, requires background knowledge. Again, pointing merely to the process of randomization does not provide justification. We need additional justification to show that our sample is similar to the real world – which is to say that our study is **externally valid**.

To conclude, there are good reasons for why one might want to randomize one's experiment – however, it is not always necessary. The importance of RCT is also sometimes exaggerated; in particular, randomization is not a guarantee for control of known or unknown background factors, and the results from an RCT cannot be generalized without further argument. In order to deal with these problems, one needs to refer to background knowledge and cannot rely solely on the procedural feature of

Data stratification: Dividing a sample into categories based on specific properties before randomizing, thus ensuring that these properties are correctly represented in the test and control groups, for instance mirroring their distribution in the population.

Known factor: A factor which you are aware of constitutes a relevant background factor.

Unknown factor: A factor which is a relevant background factor, but which you are unaware of.

External validity: An inference from an experimental system to a different target of interest is externally valid, if the conclusion holds not only for the system but also for the target, for instance an inference from a sample to a population.

randomization itself. In the light of these arguments, it is not justified to consider RCTs to be in principle better evidence than other experiments.

Part 5 - Modelling

5.1 What are models?

What are scientific models? When considering a number of examples, it becomes clear how widespread and how varied the family of scientific models really are. A simple example is a scale model, for instance a scale model of an airplane in a wind tunnel. Another example is that medical researchers and biologists use laboratory animals as models for the purpose of testing drugs that ultimately are meant to apply to humans. Yet another example is a simulation model where you program a computer to represent something. And last, we use mathematical tools – functions and equations – in order to represent certain features of the world.

In each of these cases, we are taking something to *stand in for* what we ultimately are interested in. For example, when medical researchers experiment with mice, they are not primarily interested in the animal. They are mostly interested in the animals' reaction because they are using the animal as a stand-in for a human organism. They do this on the basis of believing that a mouse, for example, has a similar digestive system as humans have, and that therefore they can test a particular substance on the mouse, and that this testing will give us clues about how humans will likely react to it.

This is one of the four features that characterize *all* these different types of models, despite the differences between them. All models are characterized by being representations**,** by containing idealizations, by being purpose-dependent, and by being manipulatable so that one can find out things about them and about the phenomena they represent. I will go through these properties in more detail.

Representation: Something is a representation of something else if it stands in for that thing, meaning we use or investigate that thing instead of what we are really interested in.

First, models are **representations**. Think, for example, of a physical model of DNA made from wood & metal. You look at this model, not simply because you are interested the artefact itself (e.g. what type of wood it is made of), but because it *stands in* for something else: real DNA. But notably, the model is not itself a piece of DNA. It is not made of the same materials as DNA, it has a different size, and, due to it being made of different materials, it has a much longer shelf life and is more stable. It was built in this particular way presumably for its pedagogic purposes – showing students how certain features of real DNA looks, while changing some of its other features so that it was easier to see, to handle, and to preserve.

Target: A target is what a model aims to represent.

Let us therefore distinguish between the model and the **target**(s) that it represents or stands in for. If a model is only a stand-in for a target that scientists are really interested in, why don't they instead investigate the target itself? After all, aren't there available experimental methods and other forms of observational methods that could be used to look for and

investigate the target? Not quite, because, for example, it is often the case that we cannot handle the target directly. Perhaps there are physical limitations such that it is in effect impossible to investigate the target directly. It might also be economically unfeasible to do such investigations. For example, we would have to build a wind tunnel for testing an actual jet. That would involve enormous resources that we typically do not have, so instead we build a scale model that we can fit into a much smaller wind tunnel.

Another reason for building models is that in many cases, we are morally or legally restricted in what we can do with or to the target. So even though we might be interested in how the human organism reacts to all kinds of substances, we are not allowed to test for such reactions directly on humans. Rather, we must instead first go through a sequence of animal testing (and animal testing, too, is restricted. We have to show that the substance has sufficient promise before we can test it on animals). Thus, we see that legal and moral restrictions constitute another reason for why we are using models instead of investigating the target itself.

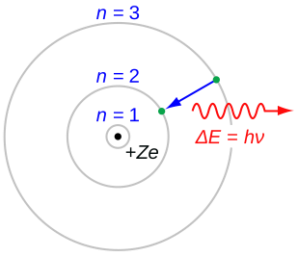
Yet another reason is that targets are often very complex. They are so complex that investigating them in all their detail would be very difficult to process cognitively. Human abilities to process information, and understand the interaction of variables, is limited. By using a model that simplifies the complexity of the target, scientists might get a better grasp of what, for example, the main causal factors operating in the system are. In this case, then, scientists are building a model not because it is not possible to investigate the target or because we are legally or morally restricted in not investigating the target directly, but because of cognitive limitations. Scientists want something simpler that helps them better process and understand such very complex systems of interaction.

Above I gave three broad reasons for why we often work with models in science rather than investigate a target system directly. These reasons were, firstly, that investigating a target system directly could be physically impossible or too expensive; i.e. economically impossible. Secondly, investigating a target directly might be legally or morally prohibited. Thirdly, investigating the target directly is in some cases cognitively too demanding, so that instead, we choose to investigate a simplifying model. These three motivations show that models are supposed to stand in for targets, but also that models must be *different* from targets – for example in their physical, economic, moral or legal features, and in the cognitive demands on users. If they weren't different, the reasons for using them would not be satisfied, and one could just as well try to investigate the targets directly.

So a model has certain properties, and a target has certain properties. Despite models being representations of targets, there are always differences between a model and its corresponding target. This is what is meant by idealization: a model is an **idealization** of a target if the model is supposed to represent that target, but model and target differ in some of their properties. We call descriptions of these differing model properties

Idealization: The fact that a model is an idealization of its target means that it does not have all the properties that the target has, and vice versa".

idealizing assumptions. So, considering the three reasons for why we choose models explained above, a common denominator is that the model is different from the target. The model might be physically smaller than its target and is therefore handled more easily, or the model might not fall under the same legal restrictions as the target, or perhaps we can cognitively process the model more easily than the target. This means that we are intentionally building or using a model in such a way that it is, with respect to at least some properties, different from the target. From this, it follows that we cannot say that whatever is the case in the model is also the case in the target, because some of the properties of the model are different from the target.



Idealization is a characteristic of models that distinguish them from theory. Consider this representation of an atom, which you probably know from high school physics class.[2] The origin of this representation is the Bohr Theory, developed about 100 years ago by Niels Bohr, who, at the time when he proposed it, had good reasons to believe that this was true. Moreover, he thought that it captured the principles that govern the behavior of atoms in the real world. Relatively soon after it was shown that, in fact, this theory was false. It had implications that were contradicted by experimental observations. Therefore, the theory was rejected. This happened somewhere in the 1920s. It may strike you as surprising that high school students still study it you still learned about this in 90 years later. Why study a theory has been proven to false? That is because you have in fact not studied it *as a theory*. Rather, your high school teacher likely presented it *as a model*. As such, it does not make the claim that it is a correct description of the world in all its consequences – like a theory would. Instead, the Bohr model comes with the caveat that it idealizes – that some of its assumptions and implications and false descriptions of actual atoms.

Despite acknowledging its various idealizations—i.e. that it will be false in certain circumstances—Bohr's account is nevertheless *useful for some purposes*. For example, it might function as a good approximation: As long as you are not interested in, let us say, the exact location of electrons in an atom, then a model as simple as this, might be good enough. Or, it might work well for the purpose of educating people about the structure of the atom.

The same cannot be said say about a theory. If a theory is shown to be false, we look for another one. You don't say that a false theory might still be useful for some purposes. Bohr's account, when it was shown to have false implications, was rejected as a theory, but kept as a model– an idealization that still works well in relation to certain purposes. This shows

[2] Picture by JabberWok from WikimediaCommons [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0/)]

that models, unlike theories, are **purpose-dependent**. We justify the use of a model with respect to the purposes of using it. If the model's idealizations do not affect the usefulness for certain purposes, then such a model can be a good representation of a target for this purpose, despite its idealizations.

Purpose dependence: A model can only be justified on the basis of how useful it is for fulfilling a certain purpose

If the quality of a model's representation depends not just on the model-target relationship, but also on the model's purpose, then there can be multiple different good models for a single target. For example, chemists use both the quantum mechanical models as well as structural models of a substance. Can we say that one is better than the other? Well, what we can say is that the two models represent their target in different ways. The quantum mechanical model is more precise in the sense that it gives you more information about the position, or the possible position, of the electrons than the structural model does. It also represents the relevant properties more accurately - it is more similar to the target than the structural model does. However, the structural model is simpler than the quantum mechanical one. It is also more transparent and is therefore easier to work with and to understand.

Lastly, the structural model it is theoretically more tractable. For this type of model, organic chemist can rely on a theory that guides their analysis. Specifically, *functional group analysis* predicts characteristic chemical reactions and behavior of a molecule, and supports systematic design of chemical synthesis, based on a functional group – a particular group of atoms in some part of the molecule. For example, by identifying a molecule by the functional -OH group as an alcohol, chemist can describe many of its properties (like solubility and boiling points compared to similar hydrocarbons) and design synthesis like e.g. esterification – to some degree independently of the other properties of the molecule. Crucially, it is much easier to apply functional group analysis to structural models than to quantum theoretical models. These models are therefore more useful in such a theoretical framework than quantum theoretic models are. The conclusion of this is that a model is purpose-dependent. We cannot say which of the two is the better model is unless we are clear about the purposes for using a model. Perhaps what we need is very high precision and very high accuracy. Or, alternatively, it is more important that we have a simple, transparent, and tractable model. These purposes will guide our choice of model.

5.2 Learning from Models

To understand how one can learn from models, we need to consider not only models' representational function, but also the way how they are manipulated. Let us start by considering the manipulation of material models. Material models are models that consist of real existing matter. The airplane scale model, for example, is made of plastic and metal. Researchers also use animals as model organisms. We can distinguish such material models from computational models and mathematical models.

Scientists do not just construct or select these models in order to study them passively. Rather, they actively engage with them and manipulate them. The scale model, for example, is placed in wind tunnel, turned at different angles, exposed to different kinds of airflow, and has its wing geometry changed. The rat is placed in a specific environment, injected with particular substances, and often gets killed and dissected. Scientists manipulate models because they want to learn about them and thereby gain new insights about the model's target. But how is that possible? To answer that question, the work of British philosopher Mary Hesse is very instructive.

Mary Hesse argued that models are not descriptions of their targets. Rather, she suggested, models function as **analogies** for their targets and thus support analogical inferences. An analogical inference starts out (i) by establishing that a model and a target are similar in certain known respects (**positive analogies**), and perhaps dissimilar in others (**negative analogies**). Then (ii) one identifies further properties $Q_1,\ldots, Q_n$ of the model, of which it is unknown whether the target has them or not (**neutral analogy**). Finally, (iii) one draws the conclusion that the target also has property Q_i, because the target is similar to the model in its positive analogies, Q_i arises from these positive analogies in the model, and the negative analogies between model and target do not disturb the relation between Q_i and the positive analogies.

Analogy: Inference by analogy is to state that a conclusion in one case applies to another case due to there being relevant similarities between the cases.

Positive analogy: Similarity between model and target.

Negative analogy: Dissimilarity between model and target.

Neutral analogy: Property of model, whose occurrence in target is not known.

Consider this example. The positive analogies between rats and humans include similar basic physiology, similar organs, and similar hormones. The negative analogies between rats and humans include the difference in size and in lifespan between humans and rats. One neutral analogy between humans and rats is the rats' reaction to a newly synthesized drug – let's say rats are capable of ingesting it, it doesn't poison them, and it cures a particular kind of cancer, while scientists don't know how humans react to it (because they are not legally allowed to test it directly on humans, say). By analogy, one then infers from the rat model that humans will react similarly to this new drug.

Of course, such an analogical inference is not always justified. Hesse proposed three requirements that analogical inferences from models must satisfy. First, there must be some positive analogies between model and target – i.e. a similarity between model properties $\{P_M\}$ and target properties $\{P_T\}$. This is pretty obvious – no analogy without at least some similarity between model and target. Second, the neutral analogy Q_i that is supposed to be transferred from model to target must causally arise from the properties $\{P_M\}$ of the positive analogy. To revert to the above rat example, the rat's reaction to the drug, the neutral analogy, must be caused by the properties that are similar between humans and rats. If one could show that for example, the anti-cancer effect of the drug in rats is caused by its interaction with a hormone that rats and humans share, then one can be more confident that this property can be transferred to humans. Third, the causal relation between $\{P_M\}$ and Q_i , including all its preconditions, must not be part of the negative analogies. For example, if it were a negative analogy how rats and humans produced this hormone (i.e. they

produce it in very different ways), and that the drug was interacting with that particular process, then one should not be confident in the analogical inference. Justifying an analogical inference between model and target thus requires further investigation: into what the positive and negative analogies between model and target are, and into the relation between neutral, positive and negative analogies in the model itself. Only if one knows sufficiently much about these relations can one be confident in the analogical inference. Sometimes considerations of background theory in combination with already established positive analogies provide that knowledge. But in most cases, separate investigation involving extensive model manipulations are required – experiments on the model that can either support or reject the analogical inferences. Neutral analogies between models and targets thus guide the way to scientific advancement.

As an example, consider the Michelson-Morley experiment that was discussed in the lecture on experiments. Up until the late 19th century, most physicists thought of light as a wave. They justified this belief by the many positive analogies between light, and water, and sound waves, such as that light produces a diffraction pattern when encountering an obstacle or slit, just like water and sound waves. Seeing water or sound waves as a model for light, these physicists then inferred a neutral analogy, namely that light, like other waves, requires a medium to travel in. This medium they called the luminiferous ether. The Michelson-Morley experiment is a consequence of this modeling attempt. The experiment eventually showed that this neutral analogy was in fact a negative analogy. This discovery eventually led more and more people to replace the wave model with newer, more informative models of light.

Thus, neutral analogies are based on model properties that we discover after, or at least separately from, those processes that, through their positive analogy to the target, helped established a model in the first place. But how do we discover these model properties? Well, to do this we need to manipulate the models. This requires that the models enable the right kind of manipulations. Not all models do that. It does not make much sense to inject a scale model with a drug, for example, or to put a lab rat in a wind tunnel. Nothing interesting can be learned from such manipulations. Thus, we must choose our material models not just with respect to the positive analogies that they offer to the desired target, but also with respect to the interesting ways we can manipulate them and learn about new model properties that provide the material for new neutral analogies. Pretty much the same reasoning can be applied also to computational and mathematical models.

For example, consider Thomas Schelling's famous urban segregation model. This model started out as a chessboard and two sets of tokens. In the first stage, tokens are distributed randomly across the board, leaving some spaces empty. In the second stage, tokens move from their original place if more than 1/2 of its neighbors are of a different kind. They move to any place on the board that has more than 1/2 of its neighbor of the same kind. This rule, iterated many times, generates stable clusters of the same tokens. Schelling suggested that this could be used as a model of

segregation in US cities, based on only a few positive analogies, in particular the spatial distribution of households and neighborhoods, the type differences between households in terms of, for example, an ethnic property, and the motivation of individual households not to be in the minority in their neighborhoods. From this, Schelling suggested a neutral analogy. As clustering occurs in the model based only on the simple moving rule, so urban segregation occurs in US cities without explicitly racist preferences necessarily being the cause.

Whatever the merit of this inference, which many social scientists indeed contest, Schelling's model shows how the neutral analogies between models and targets can guide and advance science. Schelling's model helped in the investigation of what might cause segregation. Furthermore, Schelling's many ways of manipulating the model also showed that the results could be produced from many different initial conditions, with different grids, different neighborhood structures, different proportions of tokens, and so on. Finally, Schelling's model also offered potential limits of this analogy. It exclusively focuses on the interaction between individuals and allows no role for collective agencies or government institutions. Computational and mathematical models thus shape the kind of neutral analogy inference we can make with them. Like material models, one must thus choose mathematical and computational models with respect to the modes of manipulation they afford.

If we manipulate models to learn from them, then it seems that models are similar to experiments. After all, we set variables and parameters in a model in a similar way as we would establish experimental control of background variables in an experiment. We also manipulate the model in ways akin to experimental intervention, and we are interested in observing what differences our manipulations make to the model result, similar to experimental observation of effect of intervention.

And yet there are important differences between models and experiments. To put it briefly, modeling and experimenting differ with regards to the sources and types of the most troubling errors. In experiments, the most troubling errors affect internal validity. For modeling, in contrast, the most troubling errors concern whether the relevant analogies between models and targets hold. Internal validity poses serious issues for experimenters, namely whether the inferences from experimental observations back to the experimental system are indeed valid. For this to be the case, experimenters need to carefully design and control their experiments. Now consider a model like Schelling's. Here, the modelers set the initial conditions themselves. Determining the model's parameters and programming, there are arithmetic operations between all variables. With the exception of programming errors, we can generally trust that the model results are indeed generated by those parameters and variables. The modeler is quite certain of this because they created the model themselves. Thus, at least for mathematical and computational models, internal validity issues are less of a worry for models than for experiments.

For modeling, in contrast, the main source of error is whether the relevant analogies between models and targets hold. This is often not a problem for experiments, in particular for those that are performed on the target themselves. If one investigates the target by performing an experiment on it directly, and one is confident about the internal validity of one's inferences, then one knows that the inference is applied to the target. In contrast, when one models a target and learns about the model by manipulating it, one draws on neutral analogies between model and target to apply what one learned about the model to the target. But these neutral analogies are just hypotheses. One always needs to justify why the model result should be believed to also be applicable to the target, and this is often difficult and prone to error.

To conclude, I have argued that model manipulation is an important characteristic of modeling. This allows discovering neutral analogies to the target, which often guides scientific advances, yet these neutral analogies must be justified, which is often difficult and error-prone, and this is a distinguishing factor between modeling and experimenting.

5.3 Creating Good Models – Model Virtues

The examples of scientific modelling I have discussed so far all quite diverse – they aim at different purposes, operate in different contexts and have different levels of support. If one wants to judge the quality of a model in all this diversity, a single evaluative principle would not work. Instead, I will offer a list of potential criteria, which I call "**epistemic virtues" of models**. From such a list, it will however become clear that there are tradeoffs between some of those properties. Increasing one of these virtues will necessarily result in the decrease in another. Thus, we cannot maximize all virtues at once. Rather, we have to decide which purpose we want to satisfy by using a particular model. Such considerations also translate into which of the virtues we should focus on.

Epistemic virtues of models: Types of properties of models that determine the quality of a model, given a specific purpose.

Let us start with **similarity**. Here the idea is that we determine what a good representation is by determining its level or amount of similarity to the target. The more similar a model is to the actual object that it is supposed to represent, the better the representation is. This, however, implies a rather narrow understanding of the relationship of good representation and similarity.

Similarity: The degree to which the model has several properties which the target also has (there are multiple positive analogies) that are relevant for the modelling purpose.

One suggestion, and one I think it is something that people often propose as the relevant quality criteria, is to consider goodness of representation in the following way: *a model M is a good representation of a target X, if and only if, M is similar to X with respect to properties* $\{P_i\}$*, and to a specified degree* d_i. Does this definition of "good representation" allow us to exclude some models as not being good representations of their targets? To address that question, think of any target, and think of anything that could be a representation of that target. In light of this definition of what a good representation is, given only in terms of similarity and nothing else, would we then be able to exclude some models as bad representations of any such target?

The definition that I gave you basically reduces the question of good representation to the question of similarity, i.e. a good representation has a high level of similarity. I some ways this is a convenient description because as soon as you have solved the problem of how to measure similarity between model and object, you could then compare all models with respect to their quality.

However, contrary to first appearances, that would be a very difficult thing to do - since the list of properties, according to which we are making similarity comparisons, is not limited. It does not state specific properties, nor which properties are be relevant. When I stretch out my hand in a particular way and combine that movement with a particular sound, my children will know that those properties represent an airplane. However, that should not be sufficient to say that my actions constitute a good representation of an airplane – at least not for scientific purposes. But why not?

Surely, we want similarity with respect to some specific properties. We are fine with some properties differing between model and target. What we ultimately need to select then, is *which* properties that are relevant, and therefore need to be similar. We need similarity in those properties that are *relevant given our purposes*. For example, when we put a scale model into a wind tunnel, we want the scale model to be similar in its wing geometry, for example, to the target airplane. Because we are quite confident that wing geometry is relevant for aerodynamics. The color or the cabin seating inside the airplane, however, is quite plausibly not relevant for that purpose. Therefore, the scale model need not be similar in this regard to the airplane. We can now see that similarity itself is not the recipe for a good representation. The above quality criterion is thus incomplete. What matters is only similarity with respect to those properties that are relevant for your purpose matter for goodness of representation. Relevance is determined mainly by purpose but unfortunately, modelers are often unsure whether a certain property is relevant or not for a particular purpose.

One way to deal with this uncertainty is to investigate the **robustness** of the model. This is easiest illustrated with a mathematical model. Mathematical models consist of a number of assumptions, from which a model result is deduced. Often, it is not trivial to determine whether a particular assumption affects the result or not. This is what robustness tests. We say that a model result is robust with respect to some assumption if changing this assumption (keeping other assumptions fixed) does not change the model result. If a model result is robust with respect to some assumption, then we can be quite confident that the specification of this assumption is not relevant for our purpose. Note, however, that this does not necessarily mean that we can remove that assumption from our model all together. Some assumptions, in some form, must be there so that we can calculate the model result. Robustness only shows that the specific form of this assumption does not matter. Model robustness in this sense is thus a very useful model virtue.

Robustness: A model is robust with respect to an assumption if changing this assumption does not change the model result.

Another model virtue is precision. Scientists use the term precision in many different ways. Here, I want to focus on **parameter precision**. On this meaning, a model is more precise than another model if the parameters specification of the first model imply the parameter specifications of the second. Consider the following examples. Model 1 describes the rate of change only as a function of N: $M_1=f(N)$. Model 2, in contrast, describes this as a linear function of N: $M_2=a+b*N$. Clearly, the description of M2 implies the description of M_1. If something is a linear function of N, then it is also a function of N. Therefore, M_2 is more precise than M_1 in its parameter. Parameter precision is thus a property of the model alone, not of the relation between model and target. Precision of a model is a virtue because the more precise a model, the more accurate it might *possibly* be. If the rate of change in the target is indeed a linear function of N, then the more precise model, M_2, will be more accurate than the less precise model M_1. However, precision only offers the *potential* for high accuracy and does not guarantee it. Instead, if the rate of change in the target was, for example, non-linear, then the less precise model, M_1, will be more accurate than the more precise model, M_2.

Parameter precision: One model has higher parameter precision than another model, if the specifications of the parameters of the first model implies the parameter specifications of the second.

Another important virtue of models is **simplicity**. Simplicity is quite an intuitive feature. A map, for example, is simpler than another map if it contains less detail than the first map. The map of the London Underground is an excellent example. Through numerous iterations, its makers have honed the map to only those details that are relevant for the purpose of assisting traveling the city underground. Details like distance between metro stations, curvature of the rail lines, locations of streets and monuments are all omitted. This makes the map easier to use and helps to quickly and correctly apply it for its intended purpose. However, these simplifications simultaneously make the map useless for other purposes, such as if, for example, hapless tourists try to use it to find their way around the city on foot.

Simplicity: One model is simpler than another model if it contains less parameters than another model.

A somewhat related virtue is the **tractability** of model. A model is tractable with respect to some general set of rules if the relevant model result can be obtained by applying these principles to the model. For example, a mathematical model whose result can be calculated from its assumptions using calculus is tractable, because its result can be determined using a deductively valid set of inference rules. In contrast, many nonlinear differential equations do not have an exact solution, but can only be approximated using simulation techniques. Models that can be solved using mathematical proof techniques are called *analytically* tractable to distinguish them from models whose results can only be approximated through numerical simulation methods. Tractability is a desirable virtue because it means that we actually have a powerful method at hand to analyze and solve such a model. A special case of tractability is *theoretical* tractability, in this case, the general rules of theoretical principles. The simple illustration here is that of representing a chemical compound either with a structural formula or with a quantum mechanical model. Only the former allows for the application of functional group classification. Thus, structural formulae are theoretically tractable, while quantum models are not. Consequently, while quantum models might be

Tractability: A model is *computationally tractable* only if its result can be computed in polynomial time. A model is *analytically tractable* only if the model result can be obtained through valid deduction, from the model assumptions alone. A model is *theoretically tractable* only if the model is either computationally or analytically tractable, and some of the necessary computational or deductive steps are justified with reference to a background theory.

more accurate, many chemists nevertheless prefer structural formulae. Their tractability through group classification theory often helps chemists to predict the compound's chemical behavior.

Here is an example of tractability is from the field of organic chemistry. Organic synthesis can be pre-theoretically described as some organic substances, when combined, forming new components. That description is plausible, but doesn't give an unambiguous indication which substances will combine, unless their respective combining power are represented in a particular form and are assumed to follow a general principle. Functional group analysis provides both. First, it assumes that the atomic bond structure of compound molecules is represented in a particular way (as so-called Lewis structures). Second, it identifies specific bond structures within compound molecules as responsible for the characteristic chemical reactions of those molecules (so-called functional groups). Lewis-structure models of organic substances thus allow the application of functional group analysis, and therefore the prediction whether these substances will synthesize or not. Models that satisfy the particular form required by the theory, and that therefore can be solved by applying the theoretical principle, we call theoretically tractable.

Here is another example of tractability, from the field of economics. Consumer choice can be pre-theoretically described as consumers choosing to buy goods according to availability, their desires and their means. That description is plausible, but doesn't give an unambiguous indication which goods a consumer will buy from a given set of available goods bundles and a given budget, unless her desires are (1) represented in a particular form and (2) assumed to obey a general principle. Standard microeconomic consumer choice theory provides both. First, it assumes that consumers have what is called a *twice differentiable utility function* over *the goods bundle space*, as well as a budget set constraining goods available to her. Second, it assumes that all consumers maximize the satisfaction of their desires. A model of a consumer that represents the consumer's desires in this way can be unambiguously solved with reference to this theory: the theory uses optimization under constraints as the algorithm to find all local maxima in the model's utility function. Models that satisfy the particular form required by the theory, and that therefore can be solved by applying the theoretical principle, are called theoretically tractable.

The last model virtue I want to discuss here is model **transparency**. A model is epistemically transparent if the model user is cognitively capable of understanding how the model result is produced. A competent physicist or economist, for example, is cognitively capable to understand how a theorem in her discipline is deduced from the assumption of the mathematical model. Even the most competent neurologist, however, will not be able to cognitively process how a state-of-the-art brain simulation produces its result. Such a simulation involves far too many complications of far too many interactions between variables to be cognitively processable. Transparency is desirable because it allows checking the

Transparency: A model is epistemically transparent if the model user is cognitively capable of understanding how the model result is produced.

correctness of a result. Every competent scientist can check a proof for correctness. Yet for simulation results, she needs to trust the system that performs the computations. This is particularly the case with those scientists who employ simulation platforms that others have programmed for them.

To summarize, I have sketched a number of *epistemic virtues* for models. Everything else being equal, each of them is desirable for a model to have. But as it turns out, models often cannot have all virtues at the same time. I already mentioned cases where an increase in precision, for example, might lead to a decrease in similarity. It also seems easy to imagine cases where an increase in precision might lead to a decrease in transparency or an increase in similarity to a decrease in simplicity. Model virtues sometimes trade off on each other and cannot always be maximized at the same time. This has important implications for modeling methodology. When choosing between different models, the first step is to make explicit what purpose you choose this model for. That purpose, then, will help determine which virtues are most important, how similar, in what features, how precise, how simple, and so on. The method choice proceeds by finding the best trade-off for one's purpose, optimizing model virtues as far as this trade-off allows.

5.4 Two Modelling Strategies

There are many different strategies how one can construct a good model – depending on one's modelling purpose, one will prefer maximizing some epistemic model virtues at the costs of others. In this section, I will discuss two modelling strategies: one that treats **models as mirrors** of the world, and another that treats models as isolations of particular factors in the world. Both of these strategies, as described here, are simplifying caricatures of real modelling strategies – and anyway, there are many more strategies than can be discussed here.

Mirror models: A model with high similarity to target.

It is a widely held view that our models should be as similar as possible to the target – i.e. mirror the target. Take for example simulations of epidemics. The basis of one such simulation is all the available demographic information about a real city, Portland, Oregon. Portland have about 1.8 million inhabitants, and all the available information about theses 1.8 million people are included in the simulation – where they live, how big their household is, what kind of occupation they have, whether their kids are going to school, daycare, et cetera. This information might enable us to predict how the inhabitants potentially mix with other people who might already be infected, in the case of an epidemic. This then is the basis for simulating this epidemic.

Rhetorically, such models are sometimes presented as genuine substitutes for the real system. Such models are so rich in detail, and mirror its target so accurately, that we can draw conclusions from the model to the real system without any further justification. So, for example, obviously we cannot experiment with respect to epidemics, but we can use the simulation as the basis for trying out, for example, different vaccine policies. And then we see how in the simulation things develop, but

because the simulation is a mirror of the world, the vaccines strategy that works best in our simulation will then be the strategy that we recommend for real use.

Something very similar happens in engineering when using Finite Element Analysis (FEA). Students sometimes report that they run an experiment, but then it turns out that they are actually running a simulation, and they are just taking the representation of the car body or the bridge as being a mirror of these real objects so that the simulation is then taken as a one-to-one stand-in for these real objects. But if you look more closely how the Finite Element Analysis is performed, then you see that there are numerous steps that require a lot of idealizing assumptions.

In the first place, when you start with a particular object that you want to model, and that you want to include in your simulation, you have to build a physical model of this object, and that already simplifies some of the properties of the real object. And on top of that, you are then building your Finite Element model. And here you are performing further simplifications. For example, one important one is to choose the mesh resolution. So how fine-grained do you want the mesh to impose? These days you could even choose to have the mesh be heterogeneous. In some places, you might want to have it more fine-grained than in others, but again, these are decisions you are making, all right. You are then clearly idealizing some of the features even of the physical model. And then, on top of that, you are now imposing various calculation rules. So how is this FEA model to be analyzed? What are the right ways how to now perform the calculation? And only then do you arrive at the result. So all in all, there are at least three steps, with multiple decisions about which simplifications and idealizations to allow, when constructing an FEA. There is nothing wrong with that; this is, after all, what modeling is about. But it raises the question to what extent we can actually say that the model or the simulation is a stand-in, a one-to-one mirror of the actual object, because we do, after all, include some idealizations, some of which are potentially quite relevant.

Similarly with the simulations that I mentioned at the beginning of this section – when one looks at them in more detail, one also discovers plenty of simplification decisions - assumptions that are quite likely to have effects on the result - that put this models-as-mirrors rhetoric in doubt. For example, in the epidemic model, you find that even though people are very precise about all the data that they have, it turns out they are not precise enough. The epidemic they are simulating is smallpox. In order to be infected with smallpox you need to be at a three to four meters distance to someone infected. But it turns out that the model builders did not have information about occupancy finer than a city block. A city block is considerably larger than four meters, however. You are likely to share a city block in Stockholm, for example, with people who you never have met. The simulation assumes they are mixing, however. That is clearly a strongly idealizing assumption, since, again, many of them might never have met. While it might be reasonable to assume such random mixing of inhabitants, since we do not have sufficiently fine-grained information

available, we can easily imagine how such an assumption may have a large impact on the conclusions one draws about how an epidemic might move through a city.

Considering models as mirrors is an idea that is hardly ever realized in actuality, I believe. We are clearly moving, in comparison to many other models, towards models that are very similar to a target, and they are very precise, but we are buying this by compromising on simplicity, on transparency, and on tractability. Furthermore, even though we reach high similarity, we are not reaching sufficiently high similarity to avoid the issue of external validity altogether. We will still need to argue that those assumptions are in fact harmless for our purposes. You cannot get around this. You choose, for example, a particular mesh resolution in the FE simulation because that is enough for your purposes. Clearly, if you were looking at using that model for some other purposes that measure of resolution might not have be enough, and therefore you need to make this argument, even though you have reached a high level of similarity.

Let us now consider the alternative idea that models are best conceived of as **isolations** of particular features of a complex world. This view is based on the question of how a model can be similar to its target, and be simple at the same time. Remember that we said that there might be some tension between these epistemic virtues. The idea is that models could be used to represent only some of the factors operating in some particular target system. That is, the system would not be represented in its entirety by the model, but certain factors could be represented in isolation from others.

Isolation model: A model with high simplicity.

In the early-mid 19th century, Cayley proposed that we could separate the three components of a successful airplane in the following way by arguing that we have to figure out how to obtain lift, how to provide propulsion, and how to provide control. The Wright brothers followed Cayley's division closely. They developed a separate model for measuring lift – a contraption that does not look like anything like an airplane – and they investigated that problem with this model, placing it in a wind tunnel and testing the lift properties of different foils. They also developed models of different propellers, and tested how they performed in a wind tunnel, thus separately investigating propulsion. Thirdly, they constructed gliders (much smaller than the airplane aimed at carrying a human), on which they modelled different systems of control, and they tested these under actual wind conditions on a North Carolina beach.

Arguably, it was this isolating strategy that set the Wright brothers apart from their competitors. Some tried to build an airplane on the full scale from the start, which turned out to be a huge failure that never really got off the ground. Others tried to build a full airplane, but first as a scale model. They were able to show that as a scale model, it worked. However, when they then scaled it up to full-size, and launched it from a boat on the Potomac, it went straight into the river and did not really do any flying on its way there. What we see with the case of the Wright brothers instead, is that they first developed each of these components separately. And they did so by modeling them each—by using extensive isolating models. And

only at the very late stage did they then synthesize these three components into one aircraft that they then were able to show could, in fact, fly.

What is, of course, important for this is that the system that you investigated is, in fact, separable in this way. The Wright brothers were lucky. They worked with Cayley's hypothesis that such a separation could be achieved. If there had been interaction effects, then they could not have simply synthesized each of these three components in a similar way.

To conclude this section, I have discussed the nature of models and their four characteristic features. I then discussed a number of epistemic virtues of models, each of which might improve a model, but where not all virtues can be optimized simultaneously. Last, I illustrated the complex choice of how to construct one's model by discussing two modelling strategies.

Part 6 - Statistics

6.1 Lying with Statistics

Scientific knowledge is largely empirical. That is, they can only justified through inductive inferences from empirical data. This makes methods of summarizing data and making inferences from data central to science. As I discussed in the "Scientific Inferences" chapter, such inference rules might proceed purely qualitatively, or alternatively involve various quantifications. Statistics provides such quantitative methods for data description and inference. Consequently, statistical methods have enormous importance for science. Yet, in some quarters, statistics has received a rather dubious image. "There are three kinds of lies—lies, damned lies, and statistics", Mark Twain famously quipped. For those not convinced by sayings of literary men, it might be noteworthy that one of the most widely read statistics book of the 20th century is called *How to Lie With Statistics*.

Lying with statistics: The intentional misapplication of statistical tools.

What could these authors mean by **lying with statistics**? A lie is an intentionally false statement. Intention is the key word with regards to lying. That is, one does not lie if one accidentally says something wrong. One must aim to make a false statement. In a broad sense, to lie with statistics just means to intentionally make false statements that in some way involves statistics. A researcher who writes down a bunch of numbers, falsely claiming that they are the result of an experiment, who then summarizes this "data" with the help of statistical methods is "lying with statistics" in this broad sense – although here the statistical methods are only the messenger, not the originator of the falsehood. In contrast to this, I want to focus on cases where the application of the statistical method itself is the culprit: the data is obtained from genuine observations of real processes and of high quality, and the statistical concepts and methods that are used are found in standard textbooks. Yet, when these methods and concepts are applied to the data, they yield claims that are false, or at least what most people who understand them would say are false.

How is that possible? The short answer, which I will discuss in more detail in this chapter, is that statistics offers many different methods, and while some of these methods are appropriate for some problem, others are not. *Misapplying* a statistical method from such a statistical toolbox will give false results. While those misapplying statistics without wanting to are merely incompetent, those who *intentionally* misapply statistics are lying with statistics. So one way or another, it is important to understand the toolbox character of statistics, and be competent in choosing (and justifying!) the right method for one's problem.

For example, you might ask what the average income is in your immediate family. Let us say they earn 200,000, 91,000, 39,000, 37,000 and 25,000 euros, respectively. The mean income is then 78,400 euros, which might sound rather depressing to the majority of your family members, who earn considerably less. But the median income in your family is 39,000 euros, almost half of the mean. Yet both of these numbers are correct answers to the question about the average. Mean and median are two concepts that

statistics offer us to make the common-sense notion of average more precise. When people talk about the average income, they typically have the median in mind. That is the income of the person that is exactly in the middle when one orders the incomes of all people from highest to lowest. But in other contexts, people interpret average as the mean. Therefore, statistics has defined these two concepts (and others) which, for different contexts and different purposes, pick out what most people mean by average.

All of these concepts are valid in some contexts, so there is nothing wrong with the concepts themselves. Your family's income data is correct too. And yet, by choosing one or the other concept, we can give completely different answers to the question of “What is the average income in your family?” What is more, people can cleverly employ one concept or the other to make others believe possibly false claims. For example, if you want to frustrate your sibling, the one with the 39,000-euro income, you might say the average income in our family is 78,400 euros. If he does not know the individual incomes of the other family members, he might well think about the median and hence believe that most family members make more than twice than him, thus making him believe a false claim.

Many lies with statistics indeed are promulgated by politicians pushing their agenda, journalists inflating a news item, producers exaggerating the effectiveness of a product, or lobbyists downplaying negative impact. But scientists engage in such lying with statistics practices too, sometimes for their own purposes and sometimes in the service of those lobbyists, producers, journalists, or politicians. This is particularly obvious when it comes to making inferences with statistics. The statistical toolbox, after all, contains not only concepts and methods to describe and summarize data but also methods of making inferences from data to claims about the world. Many different useful and justifiable inference methods are offered by statistics. Significance testing, for example, helps decide whether to reject a relevant hypothesis based on a set of data. Very often one hears that a certain observation is a significant result, which purportedly allows rejecting the hypothesis in question. But whether a data set is significant in a significance test depends on a whole lot of test parameters, which can be substantially influenced by the scientists performing the test. To simply say that something is significant or not does not really mean anything until all these parameters are also disclosed. Yet the habit of speaking of significant results without such qualification persists, and many people seem to only have a very vague idea of what these influencing parameters actually are.

Under these conditions, it is quite easy to devise a test based on genuine data and a valid method that yields the result one wants rather than the result that comes closest to the truth. And if this is not enough, there are other inference methods of statistics offers, Bayesian statistics for example, or a Neyman-Pearson's hypothesis testing (more on that later). Statistics, therefore, does not offer a single algorithm that takes in one's data and spits out an unequivocal result. Rather, it offers a toolbox with many different tools that can be employed for the same purposes but that

might yield different results. Consequently, one must choose the right tool for one's purposes and justify why this tool is the right one rather than some other equally applicable one.

This might come as a surprise to those who are used to statistical programs that automatize statistical-test procedures, but it actually reinforces the important distinction between method and methodology that have been discussed earlier in the course. Statistics offers multiple methods for data description and for making inferences from data. These methods yield correct results in some contexts and for some purposes. Statistics itself typically does not determine what these contexts are however. Instead, this is the task of a **methodology of statistics**; to choose the appropriate statistical methods for one's purposes and one's contexts by justifying one's choice. Only by providing such methodological justification for one's own work, and by requiring it from others, can the misuse of statistics and the promulgation of statistical lies be stopped.

Statistical methodology: Justification of the choice between statistical methods.

6.2 Descriptive Statistics

In 2013, the European Central Bank published the results of its European Household Wealth Survey (EHWS) for all European member states. The data set was large (a sample of more than 84,000 households from 20 countries), therefore, this data was analyzed using different statistical tools. Journalists who wrote about this survey did not look at the data set, but rather only at some of these specific analyses. The one that they particularly latched onto was a summary, which showed that in comparison to most other—in fact to all other—member states, the average German household wealth was lowest.

So many journalists reported that of all the member states, Germany displayed the lowest average household wealth. The ensuing discussion based on these reports has to be understood against the background of a threatening default of some southern European governments at the time, and the establishment of some form of safety mechanism against such potential defaults. Member states were asked to dedicate substantial sums to such a potential safety mechanism, and this was politically difficult. So in the discussion, one could often hear at the time: the EHWS says that German households on average have the lowest wealth of all EU members. And nevertheless, Germany is supposed to contribute a lot, or possibly most, to such a safety mechanism.

There were some voices (but surprisingly few!) who cautioned against this argument by pointing to another statistical analysis in the EHWS, which calculated the mean household wealth. Comparing these means by country, placed Germany not at the bottom anymore, but somewhat in the middle. So those people concluded that the conclusion that Germany is the poorest country in Europe does perhaps not hold. So, a question to you: if you have a choice how to summarize this data from the ECB with different statistical average notions, which one do you think is the right one for the purpose of this discussion? Try to formulate an answer before continuing!

Let us assume that the debate about the fairness of contribution to the safety mechanism by each member state depends on the overall wealth of these countries. There are surely some complications here, regarding tax fairness, corruption, etc., but I will ignore them here. But if comparative overall wealth of a country matters for this debate, then the median is misleading because it might—for example, in very unequal distributions – give a low result. The mean, in contrast, gives overall wealth divided by the number of people living in that country. So as long as it is only overall wealth that matters for this debate, then the mean is the correct measure.

However, if we now changed the question, or the purpose—if we wanted to argue about something that measures inequality of wealth distribution in Germany—then the median would certainly play an important role in any such argument. Depending on which purpose we are pursuing, we are using different concepts of average in order to perform the calculation. Thus, our choice of method for **descriptive statistics** depends on the purpose, it depends on the kind of relevant information that we have. We cannot simply say merely that statistics gives us the tools to calculate the average. You need to argue which particular average concept is the right one. With common sense notions like average, statistics offers us a choice—a menu of different methods for how to calculate the average. Methodological considerations guide our choice as to which method should be employed.

Descriptive statistics: In descriptive statistics, one aims to display data and conclusions accurately.

6.3 Evaluating Hypotheses Statistically

Statistics is used for both descriptive and **inferential** uses. In this section, I will focus on how statistics is employed to make inferences—for, example from data to hypotheses. Recall that in simple hypothesis testing, we derive an observable consequence C from a hypothesis H and check whether C is the case. If not, we reject H. If, however, we observe C, then we accept H. Admittedly, this is an extremely simple model of hypothesis testing, and I have discussed reasons why it is often too simple, but it does work for some cases. For example, I might want to test the claim that a watery solution is acidic. I know that if it is acid then when I dip the litmus paper into it, it will *not* turn blue. When I dip the paper however, I observe that it turns blue. Thus, I reject the claim that this solution is acid.

Inferential statistics: In inferential statistics, one aims to draw a justified conclusion from data.

If this simple picture of hypothesis testing is acceptable, in some cases at least, why complicate it? Why introduce statistics to make this simple and perhaps even beautiful scheme of hypothesis testing more complicated? There are three main reasons for why one might want to use statistics in one's evaluation of hypotheses. The first reason is that some hypotheses only have **stochastic** implications. That means that these hypotheses do not say anything about the world with certainty but rather only say that things might happen with a certain probability.

Stochastic hypothesis: A hypothesis whose implications come in the form of a probability distribution

A famous example is Mendel's laws of inheritance. When crossing two plants of different genotypes, one cannot predict with certainty what genotype their offspring will have. Yet, by considering all possible combinations of alleles, one can predict with what probability the offspring will have a specific genotype. For example, when crossing

heterozygous yellow and homozygous green pea plants, one might hypothesize that the offspring will be the same as one of the parents with 50% chance. But how do we test such a hypothesis? For every individual crossing experiment, we get either a yellow or a green pea. Only when we repeat the experiment sufficiently often will we get some yellow and some green offspring. But even then, when we have repeated the experiment a few times, the distribution will not be exactly 50/50. So how often do we need to repeat it? And what is a close enough proportion in our sample that we say that the observations should not make us reject the hypothesis? To answer these questions, we need statistics.

Many hypotheses, however, are not stochastic in this sense. Instead, they imply that something is or is not the case with certainty. These are called **deterministic hypotheses**. For example, the hypothesis "Amyloid plaque is the only cause of Alzheimer's disease" has the implication that *every* observed Alzheimer's patient exhibits amyloid plaque. When you investigate a sample of people, you might make the following kinds of observations: Individuals diagnosed with Alzheimer's and who exhibit plaque, individuals not diagnosed and who do not exhibit plaque, individuals not diagnosed but who nevertheless exhibit plaque, and, finally, individuals diagnosed with Alzheimer's but who do not exhibit plaque. The first three observations are compatible with the hypothesis, but the fourth and final is not. If the disease has broken out, then according to the hypothesis the cause must be present, i.e. there must be amyloid plaque present. If we follow the simple falsification scheme of the HD method that I described in chapter 2, then making such an observation should lead us to reject the hypothesis, without ever involving statistical considerations. But is it always correct to follow such a test scheme?

Deterministic hypothesis: A hypothesis all of whose implications are certain.

One reason to be hesitant is to consider that the observation itself might be incorrect. In the current case, the observation concerns two properties of patients. First, whether they have an Alzheimer's diagnosis; and second, whether they have amyloid plaque. Alzheimer's diagnosis is based on the patient's medical history and behavior and is beset by many uncertainties. In particular, it is often difficult to distinguish Alzheimer's from other forms of age-related dementia. Amyloid plaques are protein deposits in the brain. They are detected by staining brain tissue and studying it under a microscope. Clearly, these observations can go wrong. For example, when sample tissues are not properly prepared or when the microscope is not competently handled. Consequently, there might be measurement error. An observation that is produced by measurement error, however, should not lead one to reject the hypothesis. So when testing a hypothesis, one should take such measurement errors into account. In particular, if one diagnoses an individual with Alzheimer's but does not find any plaque, this is strong falsifying evidence only if one can exclude error from these observations. The more probable it is that a measurement error occurred, on the other hand, the less strong is the falsifying power of this observation. In order to make this precise, one thus needs a **quantitative measure of measurement error** – and this is the second reasons for why one might want to use statistics in one's evaluation of hypotheses.

Quantitative measure of measurement error: The likelihood of a measurement error being made, presented on a quantitative scale.

If this reminds you of the Duhem-Quine problem, you are entirely correct. Remember: Duhem pointed out that we often deduce observable consequences from hypothesis H only with the help of additional auxiliary hypotheses AH. One such auxiliary assumption is that the measurement procedure yields a correct result. Making an observation of Alzheimer's but no plaque—that implies that either the hypothesis or some auxiliary hypothesis is wrong. Unfortunately, this is an entirely negative result, because we do not know whether H or AH is wrong. So we cannot decide what to do when observing not C – reject H? Improve the measurement process? Discount this observation? But here statistics comes to our aid, because by using statistical tools we can quantify the possibility of error in a test. For example, we have a pretty good idea *how often* we go wrong when investigating plaque in brain tissue. In particular, we quantify how often plaque is not observed, although, it is in effect there and vice versa. This yields the probability of error**.** This quantified error then helps us to determine how severe a hypothesis is tested. If we observe potentially falsifying observations and the probability that this observation is incorrect is sufficiently small, then we reject the hypothesis based on this evidence. If, however, the probability of error is not small enough, then we do not reject H. It is just too probable that the seemingly falsifying observation was indeed based on incorrect measurement. This is the core idea of all **error-based statistics**, as proposed by, for example, Ronald Fisher, Jerzy Neyman, and Egon Pearson. The simplest test procedure based on this idea is the Significance Test, which I will discuss later.

Error based statistics: Determining the probability of an observation given that a certain hypothesis is true.

But first, I want to mention a third reason for using statistics when evaluating hypotheses. So far, I only discussed cases where observations lead to either accepting or rejecting a hypothesis. These decisions to accept or reject do not mean that we accept the hypothesis as true or reject it as false with certainty. This is particularly obvious with respect to accepting a hypothesis. All scientific knowledge is fallible. We should always be ready for the possibility that some hypothesis we have accepted is not true—thus, according to Fisher and his followers, we accept hypotheses although we are not certain about their truth. However, scientists are more confident in some hypotheses than in others. Perhaps, because they have particularly strong evidence for them or because these hypotheses fit well with popular theory. We can use statistics to express such differences in **confidence in a hypothesis**. We then ask how probable is the hypothesis given the observed data rather than, should I accept or reject the hypothesis given the data. We thus design a specific probability to the hypothesis, and we use statistical tools to calculate this probability. The most prominent approach in this area is Bayesian Statistics as developed by Jimmy Savage in the 1950s.

Confidence in a hypothesis: The subjective estimation of the probability of a hypothesis.

Statistics, when applied to testing deterministic hypotheses is thus a sophistication of the test procedure. It can help in quantifying test errors and also help in quantifying confidence in hypotheses. Note that this also weakens the original test procedure in various ways by allowing seemingly confirming or contradictory observations to not actually lead to either acceptance or rejection. However, this weakening is often based on legitimate reasons as I have suggested. But such weakening also opens

possibilities of abuse—of tweaking and twisting the most sophisticated tests in ways that help researchers get the result they want without necessarily getting closer to the truth – in short, to lying with statistics. Scientists, therefore, need to make a methodological decision. For every test they run, they should ask: are statistical tools needed? And if yes, which tools will be the best? I have given three reasons for using statistics here and they already indicate when statistics should be used. But when these reasons are absent, one shouldn't mindlessly continue using statistics: it weakens the test and introduces potential for abuse.

The English epidemiologist and statistician Austin Bradford Hill expressed this well in a speech in 1965. In the speech, he recounted his early empirical studies of work-related sicknesses in Britain, studies which were deliberately devoid of any statistics. In the speech, he says, "I cannot find anywhere I thought it necessary to use a test of significance. The evidence was so clear-cut, the differences between the groups were mainly so large, the contrast between respiratory and non-respiratory causes of illness so specific that no formal tests could really contribute anything of value to the argument. So why use them?" I think this is useful advice from an eminent practitioner of science. Use statistical hypothesis tests when necessary but stay away from them otherwise

6.4 Error Statistics

I gave you some reasons for using statistics when evaluating hypotheses. One of these was that we need statistics to quantify the error in testing a hypothesis with a particular data set. In other words, we have made a bunch of observations and we want to find out whether we should reject a certain hypothesis based on the data or not. For this, we calculate how probable the observed data would be if – i.e. under the assumption that – the hypothesis was true. If it is very improbable, then we should reject the hypothesis.

Fisher's significance testing: A method of statistical hypothesis testing developed by Ronald Fisher.

This is also the basic idea behind **Fisher's significance testing**. Fisher coined the term null hypothesis, but this name is more confusing than anything for the procedure only involves testing one hypothesis. Despite views to the contrary, there are no alternative hypotheses, and there is no need to formulate the null as the claim of no effect or no difference. Rather, when designing a significance test, you should test your main hypothesis – and this might consist in the claim that there is *no* effect, or that there is *some* effect, or that there is an effect in a particular *direction*, or of a certain *magnitude*, etc. This hypothesis might be false in many different ways, so rejecting it does not tell you which alternative hypothesis is true. But translating your hypothesis into a null claim (i.e. of "no effect") and testing *that* tells you even less: for example, if your hypothesis is "X makes Y increase", and you translate it to a null of "X has no effect on Y", then rejecting the null tells you *neither* that your hypothesis is true, *nor* that it is false.

Instead, you expose your main claim to a significance test to see whether you can reject it. If not, you retain it, but you do not accept it as true and you do continue testing it with other data. Let us consider an example of

such a test procedure. You have a coin and you want to test whether your coin is fair, that is, whether it falls heads or tails with the same probability. You formulate your main hypothesis H: "this coin is fair". You can test your hypothesis by flipping the coin repeatedly and observe the frequency with which it turns up heads. That test identifies the number of heads as the **test statistic**, identifies the possible outcomes of such a test and their respective probabilities, and compares these theoretical considerations with the actual number of heads observed in an experiment. The statistically interesting part of this test asks, "How far must the observation differ the most probable outcomes, in order to reject this hypothesis?"

Test statistic: All possible outcomes of a test, and their respective probabilities.

The test proceeds in these steps. First, you specify the experiment you want to run, let us say flipping the coin 20 times. Thus, the test statistic can take any value between 0 and 20. But each value of the test statistic is realized by different numbers of outcomes of the experiment – the sequences of the 20 coin flips. So second, you specify all possible outcomes of such an experiment – for example heads in the third and in the ninth throw (TTHTTTTTHTTTTTTTTTT) and tails otherwise, or heads in the first and the last five throws, and tails otherwise (HHHHHTTTTTTTTTTHHHHH). If your hypothesis is true and the coin is fair then heads and tails have the same probability for each toss, and so each single outcome has the same probability. Overall, there are 2^{20} possible outcomes. Third, you categorize all these outcomes by how they realize the test statistic – the overall number of heads and tails. For example, there will be exactly one outcome of tails only, and 20 outcomes of one heads. Fourth, for calculate the probability of each value of the test statistic. So the probability of getting exactly 1 head, given that the coin is fair, is equal to frequency of all 1-head events in all possible outcomes: $p(\text{"1 head in 20 throws"}|H) = 20/2^{20}$. This gives you the **sampling distribution** of the test statistic, under the assumption that the main hypothesis is true.

Sampling distribution: A distribution over the possible outcomes of the test statistic.

In the next step, you perform the experiment and observe the result. Let us say you get four heads. You now wonder, "How probable is it to observe this result given that the hypothesis is true?" The sampling distribution of the test statistic gives you the answer. You sum up the probabilities of the possible outcomes that are at least as extreme as the ones you observe—that is the ones you find here in the tails. This sum value is called the p-value. If your hypothesis were true, then the probability of observing your experimental result would be 0.0012, not very large indeed. This is a good reason to reject the hypothesis. This last step is often expressed more formally by saying that the p-value is smaller than an imposed threshold, the so-called **significance level**. Such levels are set by convention and differ across disciplines. As your p-value is smaller than the conventional 0.05 significance level, you reject the claim that this is a fair coin.

p-value: The probability of observing an outcome at least as extreme as the observed outcome.

Significance level: A conventionally set level for p-values, below which the associated hypothesis should be rejected.

What, however, if you do not like the test result? Perhaps you did not want to reject your hypothesis but the test said you should, or vice versa. Unfortunately, significance testing offers many ways how to manipulate

the test for one's purposes. Such behavior is called **p-value abuse**. Let us look at some examples of how this can be done. First, you could test a null instead of your own hypothesis, and hope that rejecting the null gives some support to your hypothesis – even though it does not. For example, if your hypothesis is "X makes Y increase", you could translate it to a null of "X has no effect on Y", and test that. When you reject the null, perhaps some people might think that this somehow supports your hypothesis, even though this is not the case.

p-value abuse: Changing test setup, statistical method, or sample in order to make the p-value either higher or lower than the significance level (depending on what result is desired).

Second, as long as you operate with low-powered tests (I will discuss the power of a test later), you can always shop for another test. That is, if your first experiment rejects your hypothesis, why not try it again? You can even enlarge your sample a little (as long as the power remains low) and thus claim that your second experiment is better than the first (conveniently staying silent on the power issue). And when you get a non-rejected result for one of these sample sizes, then you report that.

Third, how about changing the way you count the test statistic? In the above example, outcomes were grouped according to the number of heads. Why not group them in intervals of two or three? That will affect the p-value, and a little bit of jockeying back and forth might get you the result you want.

Fourth, what sampling distribution do you assume? In the above example, this is given by the stochastic hypothesis, so you cannot really change that. But if the error comes from a measurement instrument, for example, there is plenty of room to assume a type of distributions that might suit your illicit purposes better.

Finally, there is always a different significance level. Conveniently, people often speak about a significant result as if there was a fixed level to which your p-value is compared. So why not adjust the bar so that your favorite result can conveniently pass under it? Perhaps nobody notices.

Obviously, one should not do any of these things. These practices might further your purposes of getting a certain test result, but they undermine the whole purpose of testing, namely to expose your scientific claims to the most severe tests and, by doing so, ultimately converge on the truth. And yet these things often are done in science. In 2015, a large group of scientists, calling themselves the Open Science Collaboration, tried to reproduce the results of a hundred psychological experiment studies published in top psychology journals. The original studies all cleared the relevant significance level (otherwise they would not have been published, presumably). But when Open Science Collaboration undertook the substantial effort to reproduce these results in new experiments, they found a completely different distribution of p-values. Most of these reproductions did not clear the original significance levels, suggesting that at least some of the original results were obtained through some of the discussed illicit methods of p-value abuse.

There has been considerable skepticism about significance testing among scientists for many years. Some editors even have banned significant testing from their journals altogether. That seems somewhat exaggerated. The potential flaws, after all, lie in the illicit use of significance testing, not in the procedure itself. Therefore, my advice is to understand the procedure better to avoid and detect misuse.

What if you want to compare two hypotheses? Fisher's significance test is not applicable then. Instead, **Neyman-Pearson hypothesis-testing** approach might be helpful here as it specifies the conditions, how to decide between alternative hypotheses. The test begins by formulating two hypotheses that are mutually exclusive and jointly exhaustive. That is, one is the negation of the other. For example, we might set the $\mathbf{H_0}$ as the claim that the coin is not fair and the $\mathbf{H_a}$, the alternative hypothesis, as the claim that it is fair. Either of them might be true, hence accepting one or the other might yield one of four possible outcomes. You either correctly accept the true H_0, or correctly reject a false H_0. Besides these two correct options, you also might commit one of the following errors: If you wrongly reject the true H_0, that is called a **type I error.** If you wrongly accept the false H_0, you have committed a so-called **type II error.**

Sometimes it is more important to avoid one type of error than the other. For example, when diagnosing dangerous infectious diseases, it is worse to commit false negatives—that is type II errors—then false positives, type I errors. Neyman-Pearson's approach allows you to include these considerations into your decision about which hypothesis to accept.

In particular, *before* performing a test, one can choose the *probabilities* of committing type I and type II errors. Of course, everyone would prefer not making any errors at all, but that is sadly not possible. Rather, one can decide to maximally decrease the probability of a type I error, but only at the cost of a higher probability of a type II error – and vice versa. That is the bad news - type II error rate depends on type I error rate. But the good news is that the trade-off relationship can be weakened. First, the bigger the size of the effect you are testing, the weaker the trade-off. Of course, the effect size is not directly under your control. But if you have chosen your hypotheses well, so that their truth depends on a strongly discernible effect, then the trade-off between type I and type II error rates will be weaker. Second, increasing sample size weakens the tradeoff. Thus, in order to get a test with an acceptable type I and type II error rates, you can increase the sample size of your observation. These features of a test are typically measured by the **power of a test**. Formally, the power is the probability of correctly rejecting a false H_0 – i.e. the inverse of the type II error rate. What is really important, though, is that the power captures the interplay of type I and type II error, depending on effect size and sample size. By appropriately designing a test, one thus has some control over what error rates one is willing to accept.

Note that choosing error rates and designing a test to achieve sufficient power is not the same as p-value abuse. Two differences are particularly important. First, choice of type I and type II error are transparent, while p-

Neyman-Pearson hypothesis testing: A method of hypothesis testing developed by Jerzy Neyman and Karl Pearson.

Null hypothesis (H_0): The negation of the test hypothesis.

Alternative hypothesis (H_a): A hypothesis that due to logical necessity has to be true if the null hypothesis is false and vice versa.

Type I error: Wrongly rejecting a true null hypothesis

Type II error: Wrongly accepting a false null hypothesis.

Power of a test: The probability of correctly rejecting a false null hypothesis.

value abusers will typically want to hide their machinations. Second, choice of type I and type II error can be justified. For example, scientists can agree that it is worse to obtain false negatives than false positive when diagnosing dangerous infectious diseases, and therefore set type I and II error rates accordingly. P-value abuse, in contrast, aims to present a test decision as well supported, even though it is not. Thus to conclude, the Fisher significance test not only differs from Neyman-Pearson's procedure in that it only tests one hypothesis, but also that its decision criterion (the significance level) is arbitrarily chosen and open to manipulation in a way that Neyman-Pearson's decision criterion is not.

6.5 Bayesian Statistics

Significance testing and the Neyman-Pearson approach cover two of the three reasons for employing statistics in hypothesis evaluation. They facilitate testing stochastic hypotheses, and they allow deciding what rate of error to accept in a test. But they do not cover the third reason: they do not facilitate determining the degree of confidence in a hypothesis. Both the Fisherian and the Neyman-Pearson approach support a decision whether to reject or not reject a hypothesis. This is a binary choice based on the test results; it does not allow to reject the hypothesis "to a certain degree". Furthermore, both Fisher and Neyman-Pearson are very clear that non-rejection of a hypothesis (and in Neyman-Pearson's case, acceptance of the alternative hypothesis) does *not* mean that one must believe this hypothesis to be true. So if one wants a hypothesis evaluation procedure that helps one determine degrees of confidence in the truth of a hypothesis, one needs to turn to an entirely different statistical framework: **Bayesian Statistics**. I call this a "framework", because there are many different Bayesian approaches that all share some basic properties. Here I will only discuss the basics, as more in-depth discussions quickly get quite technical.

Bayesian statistics: Posterior probability of a hypothesis is calculated based on the prior probabilities for this hypothesis together with the observed outcome, using Bayes' theorem.

Let me begin with the test for the fairness of a coin again, but this time in light of Bayesian statistics. We start out again by formulating hypotheses, but in the Bayesian framework, we have the possibility of formulating *as many* hypotheses as we want to. Bayesians put no restrictions on the number of alternative hypotheses that we contend. However, it is important that these are all mutually exclusive. It cannot be the case that more than one hypothesis is true at the same time.

Let us imagine a situation where you only need to distinguish two types of coins: either a coin is perfectly fair (i.e. it falls heads and tails with equal probability), or it falls tails with probability ¾ and heads with ¼. Thus you only need to consider two hypotheses, H_1: "the coin is fair" and H_2: "the coin falls heads with probability 1/4". H_1 and H_2 are mutually exclusive, but they are not complete: coins can be not fair in different ways than H_2 claims. But to keep things sufficiently simple, let us assume that in this case, they cannot.

Prior probability: The (estimated) probability of the hypothesis being true before the application of Bayes' theorem.

Once we have settled on a set of hypotheses, we determine the so-called **prior probability** of the hypotheses. Here, we explicitly do what the Fisherians deny one should ever do. Namely, we are assigning

probabilities to hypotheses and, at that, before we have done any test. We could think of this practice as a test of our intuition, because we are interpreting the probabilities of **subjective degrees of belief.** How likely do *you* think it is that this coin is fair, before you start throwing it? Perhaps you see that one side of the coin has a larger impression than the other, and thus suspect that it will more probably fall on that side. Or perhaps you think that coins generally fall more often heads than tails. Or you think that the person presenting this coin is a cheat, who would likely have a biased coin in her pocket. In all of these cases, your belief might not be based on any evidence, merely consisting in a hunch, a gut feeling. Bayesians suggest that such hunches should be included in our hypothesis evaluation: before making observations, they require you to formulate the prior probability of all the hypotheses $p(H_i)$; consisting of your subjective belief in how likely it is that H_i is true. Let us say your $p(H_1) = 0.6$, and thus $p(H_2) = 1\text{-}p(H_1) = 0.4$.

Subjective degrees of belief: The Bayesian view of what is meant by "probability" – that probability is the subjective estimation of likelihood rather than a property belonging to the world.

After determining one's prior, one determines with what experiment one wants to evaluate hypotheses H_i. Let us say we use the same one as above, throwing the coin 20 times. Considering this experiment, one can determine the different values of one's test statistic, the possible outcomes for each of these values, and the probability of the different values, *if the hypothesis H_i were true*. Note that while this is quite similar as in the Fisherian case above, but now we need to determine the sampling distribution of the test statistic $p(\#Heads|H_i)$ for *each* hypothesis H_i. Note also that this probability $p(\#Heads|H)$ expresses a probability under the counterfactual assumption (that H would be true), while the prior probability expresses the probability of H being true that you currently believe. These two need to be kept separate.

Now one can collect the data – in this case the observation of the 20 coin throws. Let us assume it came out with four heads, like above. The probability of such an observation, given that H_1 is true, is 0.0046 (same argument as in the significance testing case). But now we also need to calculate the likelihood under the alternative hypothesis. The probability of such an observation, given that H_2 is true, is 0.189. (If H_2 were true, p(tails)=3/4 for any single throw, and p(heads)=1/4. Thus the probability of any sequence with four heads would be $(¼)^4*(3/4)^{16} = 3.9*10^{-5}$, and because there are 20!/(4!*16!) = 4845 such outcomes, p("4 heads, 16 tails|H_2) = 0.189. This might help if you know basic combinatorics, but is not required knowledge for this course).

With this information, one can now calculate the **posterior probabilities** of the hypotheses. These are the rational subjective beliefs one should have, given one's priors, the insight into the likelihoods, and the new observation. The calculation proceeds with Bayes' formula for the probability of a hypothesis, given the observation O that when thrown 20 times, the coin came up heads four times and tails 16 times:

$$\boldsymbol{p(H_1|O) = \frac{p(O|H_1) * p(H_1)}{p(O|H_1) * p(H_1) + p(O|H_2) * p(H_2)} = \frac{0.0046 * 0.6}{0.0046 * 0.6 + 0.189 * 0.4} = 0.035}$$

Posterior probability: The (calculated) probability of the hypothesis being true after the application of Bayes theorem.

The basic idea is that observations should affect rational beliefs in the truth of hypotheses, but only gradually. The more observations one makes, the more one's subjective beliefs take into account how the world is. In the coin-testing case, you started with the intuition that the coin was perhaps fair – you assigned slightly higher *prior* probability to H_1 than to H_2. But *after* observing the outcome of the 20 throws, you are now much less convinced that the coin is fair – you assign a *posterior* probability to H_1 of $p(H_1|O)=0.035$. Thus, the observation O made you drastically change your beliefs –you now consider it a rather remote possibility that the coin is fair. And vice versa, we will increase the probability of the competing hypothesis.

We do not then proceed to just accepting or rejecting hypotheses, but instead, we are assigning the hypotheses different probabilities with which they are true or false. This allow us to consider many hypotheses, to think of probabilities as personal beliefs, and to assign the probabilities to hypotheses. We can see now how different an approach this is from Fisherian hypothesis testing. The Bayesian approach is also computationally more involved; we need to perform more calculations than in Fisherian hypothesis testing.

Bayesian approaches have become a lot more popular in the last 20 years or so. This might be because they are computationally rather demanding, and with recent technological developments scientists now have computational abilities available to easily apply them. But that shouldn't be the reason why we accept it more. The reason should be that we think this is the right way of doing it. And here are some reasons why we should be careful about accepting Bayesianism as a nice algorithmic routine, again, where we can insert the data and get a revised, updated probability of our hypothesis. First of all, there is the issue of **determining the priors**. As mentioned above, priors are supposed to express subjective intuitions. But does that mean that any intuition is acceptable? All agree that these subjective beliefs must satisfy the basic axioms of probability – so for example, priors should always sum up to 1. But when it comes to more substantial constraints on priors, Bayesians do not agree amongst themselves about which constraints are justified and which ones are not. One such constraint is that an individual should form her beliefs in accord with her knowledge of objective probabilities (The so-called "**Principle Principle**"). Another constraint (compatible with the first but requiring separate arguments) is that if one knows nothing about a chance event involving *n* possibilities, one should equally distribute probabilities 1/n over each of these possibilities (the so-called "**Principle of Indifference**"). Others argue that none of these principles applies, and that individuals are free to choose any prior they want.

The problem with this view is that if your and my subjective intuitions differ substantially, we might never reach a time when we both have made sufficiently many observations how the world actually is so that we can come to agree. Think about someone who is convinced that tobacco does not cause cancer. If I assume a very extreme prior, it will take so much data to rationally agree with someone who looks at the same data but has

The problem of priors: Bayesianism does not offer a clear way to determine prior probabilities**.**

The principle principle: A subject's prior probability should be assigned on the basis of objective probability, if it is known.

The principle of indifference: A subject's prior probabilities should be assigned equally to the possible outcomes, if there is no information about the objective probabilities.

a very different prior, that we probably both are dead long before that. Furthermore, if I don't want to agree with you, for whatever reason, even though I make the same observations as you do, then I can choose to start with such extreme priors that my posteriors will never converge with yours, irrespective of how many of the same observations we both make. This **problem of slow convergence** has vexed Bayesians for a long time. But still they have not been able to agree what kind of restrictions on priors to impose that would solve this issue.

The problem of slow convergence: If two subjects assign sufficiently different prior probabilities to the same hypothesis, it is possible that their respective posterior probabilities will not converge even though Bayes' theorem has been applied to large amounts of data.

In the above case, I simply presented the priors as formed without any empirical observations, and the posteriors as the ones influenced by observation. But one of the great features of Bayesianism is that it describes human as continuously updating their beliefs, thus modelling accumulation of evidence over time. You start with a hunch; you make some observation and update your intuitive belief. Then, later, you make some more observations, and you update your beliefs again…and so on. Bayesian updating never ends – people keep on making new observations and update their confidence in their hypotheses accordingly. I say that's a great advantage of Bayesianism, because it captures something important about hypothesis testing that Significance Testing and the Neyman-Pearson approach neglect: that hypothesis formulation, test and experiment design are all influenced by prior beliefs.

This advantage however also bears some dangers of misuse, in particular the **problem of old evidence**. A theory that tracks accumulation of evidence over time must carefully distinguish between evidence that already has been used for updating probabilities before, and evidence that should be used in updating now. Otherwise, one might update one's confidence based on data that you have already used before. Let us say that I am running an experiment. I tell you about it, and you adjust your probabilities accordingly. Then, some weeks later, you read about the experiment again, and you further adjust your hypothesis accordingly. This is where things go wrong. This cannot happen in the null hypothesis testing account, because you are always only deciding to accept or to reject the hypothesis based on all the evidence available at one point in time. That approach is different because it does not have the sequential aspect to it. Bayesians have answered to this objection, but I am not convinced that they actually manage to give a satisfactory answer of how to distinguish old evidence from novel evidence.

The problem of old evidence: The problem of determining what evidence that has been previously used to determine posterior probabilities.

Thirdly, there is the **problem of uncertain evidence.** The standard Bayesian account describes updating as the change of prior probabilities of a hypothesis, conditional on that some observation is believed to be true *with certainty*. But as I have argued, for many of our measurement and observation results, there is no reason to assign them probability one. The possibility of measurement error, and more generally, the fallibility of our senses, should be taken into account, thus rendering many or even most claims about empirical data less than certain. How do we allow for less than certain evidence to have an influence on a hypothesis? Bayesians have offered some accounts, but they have not been universally accepted so far.

The problem of uncertain evidence: Bayesianism does not take uncertainty about evidence into account.

A lot more ground could be covered, but I hope that my main message is clear: statistics, both descriptive and inferential statistics, offers a toolbox. These tools have their advantages, but each comes with its specific problems. It is the scientists' task – those who want to make use of the toolbox - to give reasons for why they are choosing one tool over others for their purposes.

Part 7 - Explanations and causes

7.1 Explanations as the aim of science

In a previous section, I mentioned that explanation is one of the central aims of science. It is clearly not the only aim: I have already discussed some different aims, including prediction, design and explanation. Let me now try to distinguish explanation from these other aims in order to get a little clearer about what we are talking about here.

Prediction aims at forecasting: knowing if and how a particular state or phenomenon will occur at some point in the future. For that purpose, we are providing reasons for why we should expect that phenomenon to occur or not. That's different from just a lucky guess: we are providing reasons for believing that something will happen in the future. This still might not happen. Sometimes unlikely events happen instead of the likely ones. But if we offer reasons to expect a certain event, then we are justified in making the forecast in such a way even though it might occasionally not come true.

In design, we are aiming at functioning artefacts and I will say more about that in the lecture on engineering. However, what is important here is to see that what we are really doing in engineering design is that we are providing reasons for expecting that a certain manipulation of a physical object or of a structure will lead to these artefacts satisfying certain functions.

When it comes to explanation, we are aiming at *understanding*. In comparison to forecasts and functioning artefacts, it is much more controversial what understanding is. I will spend most of this lecture trying to unpack and discuss various options on how to answer this question: what is understanding? I will start from the most prominent answer to the question, at least an answer that was prominent in most of the 20th century. It is this: we understand a phenomenon if we have reasons for expecting this phenomenon on the basis of a **natural law** or a **law-like generalization**. Under this account, there is a close connection between predicting and explaining a phenomenon. We understand if we are able to give reasons for why we should have expected the phenomenon to happen, while in prediction we give reasons for expecting that the phenomenon will happen. The main difference, according to this account, between explanations and predictions is that predictions are about the future and explanations are about the past. By identifying the reasons for expecting a phenomenon to have occurred at a previous time, you have actually given an explanation.

This account is called the **Deductive-Nomological account of explanation**. It is called that way because (1) laws of nature play a very central role in it and (2) it consists in a deduction. A Deductive Nomological explanation (DN-explanation) is a deductive argument from laws of nature combined with initial conditions from which we deduce the phenomenon to be explained. Carl Hempel was a main

Natural law: A scientific law, usually thought to capture fundamental relationships in the natural world.

Law-like generalization: A generalization based on empirical evidence which appears to capture regularities in a way similar to how a law does.

Deductive nomological account of explanation: An account of what an explanation is that is based on deduction from laws and initial conditions.

proponent of this form of explanation and this is how he described what a DN explanation does:

"… a *DN* explanation answers the question "*Why* did the explanandum-phenomenon occur?" by showing that the phenomenon resulted from certain particular circumstances, specified in $C_1,C_2,\ldots,C_k$, in accordance with the laws $L_1,L_2,\ldots,L_r$. By pointing this out, the argument shows that, given the particular circumstances and the laws in question, the occurrence of the phenomenon *was to be expected*; and it is in this sense that the explanation enables us to *understand why* the phenomenon occurred. (Hempel, C., 1965a, *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, New York: Free Press., p. 337, italics in original)"

The scheme of this looks as follows. In the first place a question: we are asking for something to be explained, called the **explanandum**, E: "Why E?" The answer to the question is "Let me show how I can deduce E from a number of laws and initial conditions". These laws or true generalizations will often be in the form of some **conditional claim**, for example "if some initial conditions hold, E will be the case generally for all phenomena of that kind". Together with this law, we have a number of initial conditions C that shows that the first part of the conditional statement were the case, and from that, we deduce the explanandum.

Explanandum: What is to be explained.

Conditional claim: A claim involving the logical operator "if", for instance of the form "If A then B". (See also section 2.)

To illustrate what is happening here, take this example. At time T, we observe the planet Mars to be in place X. Now we want to know why: "Why was Mars at the position X at time T?" We don't just want to describe what happened (we already did that when describing our observation). Instead, we want to understand why that happened. The DN account does this by using a number of relevant laws, in this case probably including Newton's laws of motion, and with a number of initial conditions, such as the mass of the sun and of Mars, their present position, orbiting velocity and so on, and from that we deduce the explanandum – i.e. the position X of Mars at time T. That is it! According to the DN account, once you have provided such an argument then you have explained. You have thus generated understanding in yourself – if you asked this question – or in the person who posed this question.

Summary: Explanation is one aim of science, explanation provides understanding, and one previously prominent account says that understanding a phenomenon is achieved through deducing it from the laws of nature.

7.2 Achieving understanding

In the previous section, I described what the deductive nomological account consists in, now I want to tickle your intuitions: is this account correct? Let me start with a counterexample. Here is a Y meter high flagpole, it is a sunny day and we have the sun at this particular angle, Θ, so we are getting a shadow of length X. Here is a question for you: Which of the variables provides an explanation of the other? We have

three variables; to keep it simple, let us keep Θ fixed, so we just have the length of the flagpole and the length of the shadow.

Clearly, using standard trigonometry, we can deduce one of the three variables if we were given the other two. So, holding Θ fixed, we can deduce X from Y, and also Y from X. This seems to satisfy the DN conditions for explanation – so if we ask, "why is the shadow X m long?" we explain it by deducing it from the pole height Y and Θ; and if we ask, "Why is the pole Y m high?" we explain it by deducing it from the shadow length X and Θ. Or?

I think this last claim is problematic. True, we can calculate Y from X and Θ. But that doesn't mean that we can *explain* why Y has this and that height with reference to X and Θ. Why is that? The mathematical equations are symmetric. However, there is an interesting productive relationship that does not seem to be symmetric in the same way. The flagpole height produces the shadow, but the shadow does not produce the flagpole height. The flagpole shadow is X meters *because* the flagpole height is Y meters, but it is incorrect to say that the flagpole height is Y meters because the shadow is X meters. Even if we could change the length of the shadow, the length of the flagpole would not change. That relationship is not symmetric: it only goes one way.

This productive relationship matters because it helps us answer question like these: if we changed one of the objects, then what would happen with the other? A number of influential philosophers over the last twenty years or so have argued that if we can answer such counterfactual questions, we have gained understanding. How does one answer these "what if things had been different"-questions? According to the manipulation account, we answer them by pointing to productive relationships: What produces changes in other objects or other events? That is what we call a **cause**. This analysis gave rise to alternative views of what understanding consists in, different from the DN account, called **causal accounts of explanation**. It argues that what is important for understanding is the ability to answer such a "what if things had been different" question.

Therefore, I would argue that the answer to our flagpole case is that we gain understanding by pointing out the height of the flagpole, because that is the cause of the shadow in these conditions. We are however not gaining the understanding of the flagpole height by pointing to the shadow, because what produced the height of flagpole was something else altogether, namely decisions made by the company that produced the flagpole or the desires of the client who ordered the flagpole. That is what caused the flagpole to have that length. According to causal accounts of explanation, one offers a genuine explanation responding the question "why is the shadow X m long?", by deducing X it from the pole height Y and Θ; but deducing Y from the shadow length X and Θ, does *not* provide a genuine explanation responding to the question "why is the pole Y m high?".

Cause: Something which produces a change in another object or in another event.

Causal account of explanation: An account of explanation where causes are used as the explanans (what provides the explanation).

With this account, you can also say there are degrees of understanding; you have *some* understanding if you, for a particular domain, are able to answer *some* counterfactual question. You might still say, "I do not know an answer to a lot of the possible counterfactual questions". The more counterfactual questions you are able to answer about a particular domain the more you have gained understanding.

Note that this is *one* alternative proposal of what understanding might be, and there certainly are others. As I said, what understanding is remains a controversial question in philosophy, and I cannot do justice here to all the different arguments in this debate. The only argument that I am making here is that the DN account is defective in at least some cases, and that the causal account of explanation captures important intuitions in those cases better than the DN account. This leaves open the question whether the causal account, in the way I am presenting it here, might need further revision or at least specification. This is what philosophers are working on today!

The flagpole example is a bit of a toy example, but you can find similar issues all over science. In science, we use mathematical equations and equations in their nature are symmetric. Consider the relationship between the pendulum length *l* and its period *T*:

$$T = 2\pi * \sqrt{\left(\frac{l}{g}\right)}$$

We can rearrange this equation to its logical equivalent:

$$l = T^2 * \frac{g}{4\pi^2}$$

Thus, we can either solve the equation for period *T* or we can solve for the pendulum's length *l*. There is no logical difference here; we are gaining information about that variable from the magnitude of the other variable. Yet, I would argue again that when we want to gain understanding, the symmetry of this equation is misleading. The length of the pendulum plus further properties produces the period of the pendulum. If we ask, "why does the pendulum have this period?", then we can point out that it has this length etc., but the other way around, if we ask "why does the pendulum have this length?" it would seem a bad answer to then say "well, because it has this period". The period does not produce the pendulum's length.

Moreover, we can test that by asking a counterfactual question: what would happen if we changed the pendulum's period? Even if we did so, maybe by placing a strong fan one side of the pendulum so that it blew further to one side, we would not expect the pendulum's length to change. Therefore, we have the asymmetry in the productive relationship; and this is what we need to take into account when we make an explanation. The DN account is not able to make that distinction. The DN account argues that it is the deductive relationship between the initial

conditions and true regularities that provides an explanation and because of this, we should be able to use these equations to explain either the period or the length of the pendulum. Thus, the DN account is not *sufficient* for explanation in all cases. It does not sufficiently distinguish what gives us understanding from what does not, because it does not distinguish the symmetry of the equations. That is a first argument against the DN account. The DN account is unable to discern what should be discerned.

Here is another argument against the DN account where it fails to discern what it should discern – namely causally relevant from causally irrelevant factors. Here is something to be explained. "Why does Mr Jones fail to get pregnant?" A reasonable question that might require an explanation. Here is a true generalization: All biological males who take birth control pills regularly fail to get pregnant. Now, in the example there are initial conditions that match these general claims: Mr Jones is in fact a biological male who has been taking his birth control pills diligently. Thus, using this generalization and the initial conditions we can deduce that Mr Jones fail to get pregnant, and according to the DN account, we have explained why this happened. However, I hope you agree to consider that strange. The birth control pills are causally irrelevant here. They do not produce the phenomenon pregnancy prevention – although in other cases they can be productive causes! The productive causes in this case are the anatomical differences between biological males and biological females. That is what actually brings about why Mr Jones fails to get pregnant, and therefore explains it. The DN account does not distinguish between causes productive in particular cases and those that are not, but rather relies solely on generalities about productive relationships – and it is true, after all, that birth control pills tend to prevent pregnancies. Because of this, the DN account is not sufficient to provide explanations in all cases.

The DN account sometimes identifies arguments as explanations although the do not produce understanding. It also sometimes denies that certain arguments are explanations, even though they produce understanding. Or, to put it in another way, the DN account is not *necessary* – something can be an explanation without being a DN-explanation. One example is **singular causal explanations**. What we often do when we give a causal explanation is that we simply describe the chain of singular events. Explanans: "Why did the vase break?" Explanandum: (1) It stood on the table at a certain height, (2) it was a hard floor below it, (3) I walked passed the table, (4) I brushed it, and (5) the vase fell down. The explanans includes no law or true generalization. Admittedly, we could build one into it: "If something is a fragile object placed at over a height of X over a hard surface, then it breaks", but really, that is not the point. It seems we are perfectly satisfied with giving these statements as explanandum, which are just sequences of particular events, without deducing the event from any particular law.

Singular causal explanation: The explanation of a singular phenomenon or event by stating particular events as causes in the explanans.

Engineers, when searching for errors in a device, are looking for an explanation: why does this device not work? Often, a good answer for this would be such a sequence of singular events, because somewhere in this sequence you would find something that makes you say, "Ah, maybe the switch here was laid in the wrong way". You don't need a regularity or cite any laws of nature, it is just enough to give that sequence and that will be a perfectly acceptable and in many cases a very useful explanation. The DN account excludes it since it says you need to have a regularity from which you deduce the explanandum and that seems to be too limited. Therefore, it does not even produce necessary conditions for a good explanation.

As I said before, many philosophers have argued that the DN account should be rejected in favour of the causal account of explanation. Explanations, according to the alternative account, consists in the provision of *contributing causes* to the phenomenon that is to be explained, that make a difference in that situation. I will come back to describing in more detail what "making a difference" means. Identifying these difference-making causes helps us to answer these "what if something had been different questions". What is important to mention here is that I am only stressing that such a causal explanation is *sufficient* for understanding. There is more lively controversy over the question over where it is also *necessary*. It might be for example that certain kinds of mathematical explanations, that arguably are not causal, still provide us with understanding. I am not saying that it is necessary that you identify difference-making factors in order to obtain understanding; however, it is sufficient. It is *one way* we can attain understanding.

With this account, we get a difference between explanation and prediction, and if you think about modelling exercises then you can see how big this difference in fact is. If we are building a model for a purely predictive purpose, we are likely to build a model in a very different way than if we are building a model for an explanatory purpose. For a predictive purpose, what is most important is that we have a relationship between input and output variables so that we are able to describe how a certain state in the future will look. For example how the weather will be tomorrow or the day after. What the exact causes are in this model is not so important – elsewhere I argued that such models might deliberately deviate from what is known about the true causal structure behind the phenomenon. In contrast, a model that purely aims at explaining something and providing understanding will instead pick out only the difference-making causes. It will not care about many of the background causes at all, and it might be a simple model because it only isolate the difference making causes. That is why many explanatory models cannot predict at all. It is helpful to understand this difference to understand why we have to build and use models in different ways for these different purposes.

7.3 The format of an explanation

So far, I have focussed on the content of explanations. Now I want to say a little more about its form. An explanation starts with a question. Why did the vase break? The explanation itself then consists of the satisfactory answer to that question. Here is a little bit of terminology: we call the question, that which identify what is to be explained, the **explanandum (plural: explananda)**, and we call that which provides the answer the **explanans (plural: explanantia)**.

We can now distinguish between different kinds of explananda. We might have a **singular explanandum**, a particular event, phenomenon or property: the vase being broke, that particular car being involved in an accident, a particular historical event. However, we can also have a **general explanandum**, which is not particular events. Take this example: the relation between the gender-gap index and the percentage of women in STEM (Science Technology, Engineering and Mathematics) professions. In the gender-gap index, the less equal men and women are treated, the lower the gender gap index, the more equal they are treated the higher the index. Finland, Norway and Sweden have a high index and there are a number of countries where the gender-gap index is very low. Nevertheless, interestingly, the lower the index the more women are graduating in STEM subjects. That is obviously not something that you can just observe as a singular phenomenon. Instead, you need to do a lot of measurements and modelling for each of the countries and you need to operationalise these measures. The phenomenon to be explained is "why is there a negative relationship between gender gap index and percentage of female STEM-graduates?" In science there are often phenomena like this that we want to explain: general explananda.

Similarly, there are different kinds of explanantia: singular and general. A **singular explanans** consists of a description of a number of particular events, phenomena or properties, like the singular causal explanation of the vase breaking that I discussed above. A **general explanans** consists in a regularity or law. You might think, "Isn't that going back to the DN account? And indeed, the deduction from laws was one of the DN hallmarks. But it wasn't the reference to regularities that was problematic about it. Rather, it was the fact that sometimes these regularities do not represent any, or not the relevant, causes needed for an explanation. The flagpole and Mr. Jones cases were examples of such regularities. In many cases, causal knowledge and productive relationships are represented though general regularities and we can use those. There is nothing wrong with causal explanation accounts using such regularities. A general explanans would make use of such causal regularity.

One important aspect of the form of an explanation is that it is **contrastive**. I have already mentioned that an explanation should point to the difference-making parts of a situation. Really, when we are asking "why did this happen?" we are asking, "Why it did happen in this way

Explanandum (explananda): What is to be explained.

Explanans (explanantia): What provides the explanation.

Singular explanandum: A singular event, phenomenon or property is to be explained.

General explanandum: A general event, phenomenon or property is to be explained.

Singular explanans: The explanation is provided by a number of particular events, phenomena and properties.

General explanans: The explanation is provided by a regularity or a law.

rather than in some other way?" or "why did this event happen rather than this other event?" Unfortunately, far too often, this contrast remains implicit in explanation; we often do not give an explicit contrast of what we are asking for. This makes it very difficult to get the answer right, because the explanans needs to respond to that difference, that contrast that we are asking for in the question. When we are asking the question "why did the vase break", the full question is implicit, for instance we might really be asking, "Why did the vase break in to fragments rather than just show fissures?" By spelling out the entire question, we make sure that the answer addresses that particular difference.

Here is a scenario that illustrates the importance of making the contrast explicit. There was an accident. A police officer and a road engineer both ask: "Why did the crash occur?" The police officer gets the answer "Because the driver was intoxicated" and she thinks that is an appropriate explanation. Let us assume that that is true too, that the driver was indeed intoxicated. The intoxication caused the crash to occur. Now we have the road engineer and he asks the same question: "Why did the crash occur?" Now he gets a different answer: "Because the curve was too tightly banked". Let us assume that is also true, because they might both be true: the driver might be intoxicated and the curve might be too tightly banked.

Both of them ask the same question but they get different answers, but they are both considering that the answer is complete. That seems to be troublesome, because if we analyse this we would have to admit that the same question would have to be given different correct explanations, and that would seem strange. What is happening here is of course that they have different contrasts in mind. The police officer asks why this crash occurred *with this driver rather with other drivers who traversed this particular curve safely*. Her focus is the personal culpability of the driver. In contrast to that, the engineer asks, "Why did the crash occur *in this road stretch, when it did not occur in the previous curve*?" His focus is on the infrastructure.

What seemed like the same question, being given different correct explanations, upon closer inspection turned out to actually be two different questions, differentiated by their respective implicit contrasts. The conclusion from this is that you need to make clear what the contrast is in your explanandum, in order then to properly address that difference in explanans.

7.4 What makes explanations powerful?

In the previous part, I have talked about the content and the form of explanation and I argued that you have to identify difference-making causes; you have to address the contrastive explanandum, for something to be an explanation. Now I want to discuss what makes an explanation better or worse: the quality criteria of an explanation, and in particular, the quality criteria for causal explanations. Here are five features that we might call **explanatory virtues**. These are not necessary characteristics of explanation – those I discussed in the first part of this lecture, when

Explanatory virtue: A property of an explanation which makes it a good explanation.

distinguishing explanation from e.g. prediction. Here instead I am asking, "If something is intended as an explanation, what makes it a good or a bad explanation?"

The first one is **accuracy**. Put very simply, a good explanation must be true. If you come up with a "just so" story – e.g. Camels have humps because they were punished by the Djinn of All Deserts for being idle, as Kipling tells it – you have not given an explanation or provided anyone with understanding. Maybe they have a *sense* understanding, but such a psychological sense of understanding might mislead. Distinguish between this feeling and the actual ability of answering a "what if things had been different"-question.

Accuracy (explanations): The explanation is true.

Note, however, that accuracy is limited. We do not want to say that an explanation has to be completely accurate. Take the example with models, which can be used to explain phenomena. By saying, "if you want to explain with a model the model needs to be true in all parts", you will exclude *all models*, since models are always idealized and they always omit something. That is what makes them *models* of something rather than the real thing. Instead, the accuracy criterion for good explanation requires that one should accurately identify the difference-making contributing causes.

But not everything is a difference-making contributing cause. If you ask, "why did this vase break in fragments rather than fissures?" you need to address rather particular features of the vase and the height between the table and the floor, but there are many other factors you can disregard. While factors such as that the vase existed at all or the gravitational pull towards the centre of the earth clearly are required for the vase falling or breaking at all, they did not make a difference to what you are asking for. Anything that does not make a difference for what you are asking for in the explanandum can be disregarded in the explanation. Accuracy only concerns these difference-making causes.

Besides accuracy, we need precision. Precision here in the context of explanation in the first place means **precision of the explanandum**. We want the contrast in the explanandum to be as precisely stated as possible. Recall the discussion of the police officer and the road engineer, we need to have them make the contrast explicit for us to understand what they are actually asking for, otherwise we might give them answers that are irrelevant even though they might be, in some sense, correct. They give them information, but not the information that they have been asking for or that would help them in their understanding.

Precision of the explanandum: The contrast featured in the explanandum is precisely stated.

Once we have precision in the explanandum, we then also need to have **precision in the explanans**: to reflect that contrast in the explanans. Identifying difference-making causes is not trivial, you need to not only provide causes that truly have been operating in the situation, but you also have to provide those causes that actually made the difference. Those have to be accurate, but they also have to be precise. Clearly, an

Precision in the explanans: The contrast featured in the explanandum is reflected in the explanans.

explanation can go wrong here: an explanation might not pick out the difference-making causes at all, for example. Or it might pick out the difference-making causes and a lot of other causes that didn't make a difference. You would then improve the explanation by throwing out all of those causes that were operating but did not make a difference to the question.

Non-sensitivity: The explanans is not sensitive to small changes in circumstances, so that the causal chain in the explanans is not too long.

The next explanatory virtue is **non-sensitivity**. Built into the notion of contributing causes is the idea that there are causal chains between events, where one event causes the next event, which causes a further event and so on. I will explain this idea of direct and indirect causes more in a section below. This chain of causes can get very long – it could be argued that this chain, strictly speaking, is as long as time itself – and the longer the part of this chain that we use in our explanation is, the more background conditions must be in place for such a contributing cause to actually be difference making. The more additional conditions you need for the causes you are citing to be a difference-maker, the more sensitive your explanation becomes.

Lenin once confessed to Maxim Gorky that he loved Beethoven's *Appassionata* – but that he avoided listening to it "it makes me want to stroke people's heads, and I have to smash those heads to bring the revolution to them". So, did the October Revolution happen because Lenin did not listen to Beethoven? Such an explanation would be extremely sensitive. A lot of background factors would have to be satisfied for this explanation to work, and particularly with historical explanations, we cannot be sure that these assumptions would really hold. This explanation is too sensitive and we should instead aim for difference making causes that do not require a lot of very particular background conditions to be in place.

Cognitive salience: The explanation is easy to understand.

Last, an explanation is an answer to a question stated by a person or a group. Therefore, the answer to the question should be responding to that person's, or that group's, cognitive abilities. An explanation should takes into account the cognitive abilities of the person or the group asking the question. This is called **cognitive salience**. Typically that means you want to keep the explanation simple. One way to keep it simple is to not include, as we already discussed, descriptions of causal factors that do not make a difference for the explanandum. Another way to keep things simple is to keep the explanation on the highest macro-level that is still feasible in order to explain.

Think of the price of a company share traded at a stock exchange. Economists often explain price levels and price changes by reference to aggregate demand and supply functions, which are typically estimated from past market data. But of course, these aggregates consist of the choices and interactions of individual traders, so it would be more accurate to perform an explanation of prices based on the explicit representations of these individual choices and interactions. We are getting more and more fine grained in our detail in how we are describing our contributing causes, and we might even get more

accurate in describing these causes but we are losing a lot of cognitive salience. So here, there is a trade-off perhaps between how detailed we describe the causes and how well others comprehend our explanation. That needs to be taken in to account too.

Summary: I have argued that causal explanations are difference-making, consist of identifying difference-making contributing causes, and that we can have at least five quality criteria, by which we make these explanations better.

7.5 What is causation?

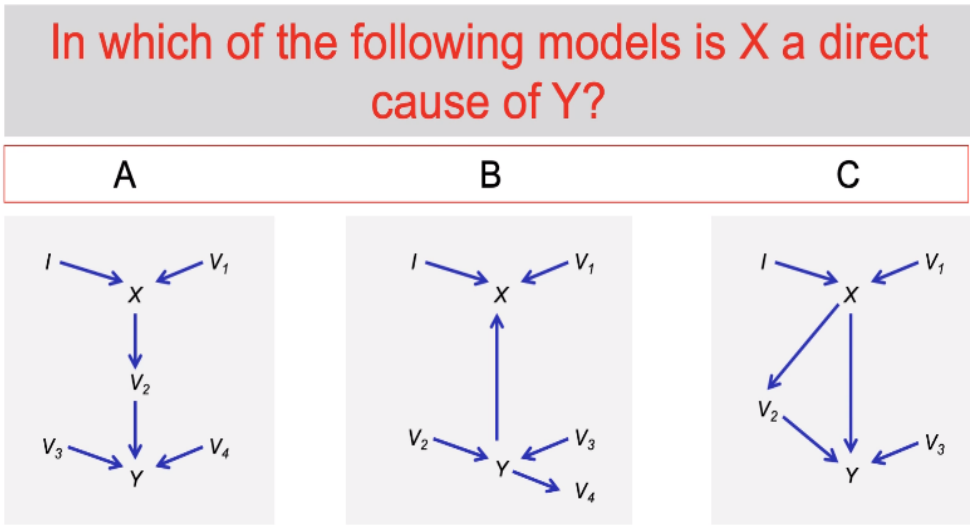
So far, I have taken the concept of cause for granted – I have only characterised it as a productive relationship between two events that is asymmetric and thus different from a (symmetric) mathematical equation. Let me now offer a more precise definition. The basic idea is the concept of a **direct cause**: we have two variables, X and Y.

X is a direct cause of Y *if and only if* there is a possible intervention on X that will change Y, when all the background conditions have been held constant.

This definition identifies the causal relationship between X and Y under the assumption of given set of background variables, and also under the assumption that there is a possible intervention on X. Therefore, we cannot say that X causes Y generally, but only against the particular background conditions that we are talking about. We often colloquially talk about causes between factors without specifying what background factors we are assuming. This approach makes it explicit that we are indexing these factors.

We can graphically represent this, we have here X, and we have the arrow that represents causal relationship where X is a cause of Y, we have a number of background variables, $V_{1...n}$ that influence either X or Y, and we have the intervention I on X. And now we are holding these background variables constant and then we are performing the intervention I on X, and then we see whether Y actually changes. Only if this intervention on X leads to a change in Y, given that all the background variables are constant, then we say that X causes Y.

Let us try this out. In which of the following models is X a direct cause of Y?



Only in (C)! Let me explain why this is. In (A), we have a causal relationship to Y, but it is not a direct cause. Instead, it is mediated by background factor V_2. As the definition says all background variables V are held constant, which in this case means that if necessary, they are

manipulated to retain their previous value. Therefore, V_2 screens off the influence of X on Y. Therefore, Y would not change even if we intervene on X.

In case (B), the influence goes the other way around, Y influences X, if we intervene on X, we will not see a change in Y. Therefore X does not cause Y, Y causes X. Finally, in (C), we do have a direct connection between X and Y. (There is also have an indirect connection which will be screened off, by holding V_2 constant). But since there is a direct influence of X on Y, an intervention on X will lead to a change in Y, thus we say X causes Y.

From the definition of a direct cause, we can now derive a definition of **contributing cause**: X is a contributing cause of Y with respect of a set of background variables if there is a causal chain, each link of which is a direct cause extending from X to Y. Thus in (A), although X is not a direct cause of Y, it is a contributing cause of Y.

Contributing cause: X is a contributing cause of Y with respect of a set of background variables if there is a causal chain, each link of which is a direct cause extending from X to Y. Thus in (A), although X is not a direct cause of Y, it is a contributing cause of Y.

Now you can see why causal explanations refer to contributing causes rather than direct causes. One can always find a sufficiently fine-grained description were we can point to some intermediate steps between whatever one phenomena or description end and our explanandum factor. If we are explaining on the social level we can find intermediate causes in the human level, in the human level we find them in the biological level, in the biological level we find them in the chemical, then in the physical level – and then, who knows. We cannot require our explanations to cite direct causes. We do not want to require a direct cause for an explanation, but rather any point in the chain, which will be enough for an explanation as long as it satisfies accuracy, difference making, non-sensitivity and cognitive salience. The longer the chain gets the longer the more sensitive the explanation will become, and we do not want it to become too sensitive.

Summary: I gave you an account of direct cause in terms of the manipulation that we perform and the reaction we observe from it, and I derived from that the notion of contributing cause. With that, we can now delve into what I think is the most pressing question: how do we learn about causes?

7.6 Learning about causes

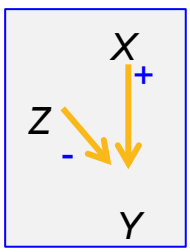
How do we learn about causes? I assume you are all aware that correlation is not the same as causation. Correlation is a measure of association between variables, while causation measures the productive influences of one variable on another. These two relations clearly are different. We already have uncovered a number of differences, one being the symmetry of correlation (if X is correlated to Y then Y is correlated to X) and the asymmetry of causation (if X produces Y, it certainly does not imply that Y produces X, and it is in fact rather unlikely that Y produces X).

There is another difference, which made causation suspicious for many philosophers of science. David Hume argued already in the 18th century that we only observe associations, correlations, we never observe causes themselves. Many people agree that we cannot directly observe causes. It is relatively straightforward to observe correlation; it is much more complicated to observe causal relationships, if we can observe them at all.

For this and other reasons, in the early 20th century, some philosophers (in particular those relating to logical empiricism) and many scientists (in particular those in physics) argued that causal language should be avoided. Speaking about causes, they maintained, means that you use words that refer to phenomena which we have difficulties observing. Instead, we should use algebraic equations (or probability calculus) and then we do not need causation. Now, I began this chapter by arguing that this is actually not true, and that in order to acquire understanding we need to go beyond the symmetry of mathematical equations and we need to actually interpret them in terms of a causal relationship.

Because causation is difficult to observe, correlation is certainly an important kind of evidence, or maybe the most important kind of evidence, for causal relationships. First, let us discuss how correlation and causation differ, in particular when we have a correlation between two variables and what this means for our conclusions about the causal relationship between two variables.

The first thing to note is that correlation between two variables in a data set is not necessary, even when there actually is a causal relationship between them. You might find that there is no correlation between two variables, X and Y. One way of how to measure the correlation is covariance, and in your dataset the covariance is zero or near zero and yet there might a causal relationship between these two variables. How is this possible? Well, there might be another cause, hidden to you, that influences Y, so that the influence of X on Y is hidden. X might have a positive productive influence on Y, but that there might be a negative influence Z on Y, maybe even exactly to the same extent that X has a positive influence on Y.



In this case, we have a causal relationship but we do not have any correlation. Therefore, it *is not necessary* to find correlation even when there is a causal relationship between variables. One way to deal with this is to try to control for Z: if we were able to break that link between Z and Y, then the causal relationship would suddenly appear in our

measurement of correlation. That is what a lot of experimenting is about, actually performing these controls in such a way that such causal relationships appear clearly to us.

Therefore, correlation is not necessary for causation and we need to work hard so we uncover these relationships. Correlation is *not sufficient* for causation either. We might have a very strong correlation between X and Y, yet that correlation might come about through very different **causal scenarios**. It might of course be that it comes about through X actually causing Y, but it might also come about in other ways, for instance by Y causing X. Just because there is a correlation, we cannot conclude what is the directionality of the productive relationship. Sometimes we might get that wrong.

Causal scenario: A description of the causal chains in a situation or in a system, such as A causing B, B causingA, or C causing both A&B.

Sometimes it might be that we have no causal relationship between X and Y whatsoever, Y does not produce X, X does not produce Y, yet X and Y are highly correlated. This is because there is some third factor C that causes both X and Y, thus making them strongly correlated. Here are some examples. Per capita candy consumption highly correlates with divorce rates. You then intervene on candy consumption and then nothing happens with respect to divorce rates. Why is that? Because there is a **common cause** C: people who are younger eat more candy and they get divorced more often. In a population, you find such a correlation between divorce rates and candy consumption unless you are screening for age. Once you screen for age, separate age, then you of course do not.

Common cause: Two phenomena are caused by the same phenomenon.

Here is another example: hormone replacement therapy, which is often done do counteract some of the negative consequences of menopause. It was found that there was a strong negative correlation between hormone replacement therapy and being affected by heart disease. Given that the dataset was large and that the correlation was strong, it was then suggested that hormone replacement therapy could be used to treat heart disease. Well, later it turned out that there is a correlation but there is no causal relationship. There is a background common cause and that is socio-economic status: women who are wealthier and more educated are more likely to get hormone replacement therapy during the onset of menopause and they are also at the same time less effected by heart disease. However, there is no relationship between the two variables; instead, there is a common cause.

Then there are the more mysterious cases when we have a correlation between X and Y but we do not find any causal narrative for it at all. One example is per capita consumption of mozzarella cheese and civil engineering doctorates awarded. That has a correlation in US data from 2000-2010 somewhere in the 99% range. Another example is that US spending on science, space and technology correlates strongly with suicides by hanging, strangulation and suffocation. We do not know why this is. One possibility is that this is no common cause; this is a spurious correlation. The reason for the covariance might simply be that at least for in the first decade of the 20th century these were reliably increasing, both of them, it is the increasing at similar rates but for unrelated reasons. If

you went out looking for correlations, using the vast amount of data and the amazing computer power we have today, you would find a number of these correlations. “Big data” – this not very scientific but popular term – does a lot of that stuff. We can crunch many numbers and we get many correlations. Well, what for? Maybe to amuse your students with such stories, but unless these searches are done with a particular purpose in mind, based on a previous understanding of possible causal relationships, they will not yield any valuable information about actual causal relationships.

What is important here to see is that correlation is not sufficient for causal relationships. Instead, for any given correlation between two variables, there are usually many alternative causal scenarios that might match and explain that correlation. Therefore, we need to now have separate evidence that tells us which of these causal scenarios is the right one. So we have the correlation, the correlation does not tell us which of the causal scenarios is the right one, we need something more.

Here are two strategies to actually achieve this. The first one, you already know: Mill’s method of difference. You control all background variables influencing X and Y and you intervene on the hypothesised cause to see if it results in a difference. If Y is in fact caused by X and you intervene on X, keeping all the background factors constant, you will see that X changes. In all of the other cases, there will be no effect. If we intervene on X but Y causes X, we will not see any difference in Y. In a common cause scenario where we intervene on X and we keep all background factors constant, including the common cause, there will be no difference on Y. Thus, Mill’s method of difference gives a way to learn about causal relationships but it requires that we perform experiments.

Often experiments are not possible, feasible or legally permitted – instead of intervening, we can only observe. What do we do with all the observational studies data; can we get from that to causal relationships? The answer is yes, and one of the most important ways is **instrumental variable analysis**. You have is correlation between X and Y. Now you find another variable, that you know produces X but has no causal relationship to Y itself (besides perhaps through X). Then you can use any correlation that exists between Z and Y in order to determine whether X causes Y. Here is an example.

Instrumental variable analysis: Determining causal relationships from data.

There is a strong correlation between smoking and deteriorating health. Scientists knew this for many decades but it took an enormously long time before policymakers started to act. Why? Because it was just a correlation and not a causal relationship. We cannot make people smoke and then see if their health declines, so experimental studies are out of the question. The observed correlation might be the result of a wide variety of causal scenarios. It could be because (i) smoking causes poor health, but it could also be that (ii) those who have bad health are somehow drawn to smoking. Alternatively, there might be a (iii) common cause: maybe people who are depressed are more likely to pick up smoking but those who are depressed also tend to suffer from worse health. Fisher, the

famous statistician who himself was an avid smoker, argued that perhaps there was a gene that caused people to suffer from poor health and also caused them to desire to smoke. Finally, (iv) it might just be something that goes in parallel over time, people get worse in health. Only the first scenario would justify a policy of smoke prevention for the benefit of smokers' health. As long as policymakers were unsure about the scenario (and there were many vested interests in keeping up that uncertainty), they had no good reason to act.

Here is an instrumental variable approach on how to solve this. Smoking has been taxed for a long time, like other luxury goods, not for reasons of health policy but in order to generate revenue for the government. Governments in many countries have repeatedly increased tobacco taxes, so there is a wealth of observational data about this. In this data, we might discover that the introduction or increase of cigarette taxes actually changes peoples smoking behaviour. Cigarette taxes are therefore causing a change in smoking, but cigarette taxes should not cause changes in health rates except through fewer people smoking. Taxes thus are the instrumental variable causing a change in smoking (X) that we know (from theoretical considerations) do not cause changes in health (Y).

If we now observe, say a year after tobacco taxes were increased, that cigarette consumption (X) went down and smokers' health (Y) went up (and we can exclude other factors affecting Y), then we have a good reason to exclude all but the "smoking causes poor health" scenario. We know that X decreased because of the tax hike, hence scenario (ii) can be excluded. We know that X does not affect Y (other than perhaps through X), so it cannot be a common cause, as in scenario (iii). And scenario (iv) cannot be correct either, because the tax hike just created a sudden drop in X, out of its long-term fluctuations. Thus by excluding scenarios (ii) to (iv), only (i) remains as an explanation of the correlation between X and Y: the instrumental variable analysis helped us to identify smoking as a cause of bad health, based alone on observational data. (Note that the data is not generated by any kind of experiment, as there is no controlled intervention; nor can we consider the observational study a natural experiment – all smokers are taxed, so we cannot divide the population into those smokers affected and those unaffected by the tax).

This is really just the very basis of a statistical technique of picking out causes from observational studies, but I think it is important to see that we do not necessarily need experiments to identify causes; we also have means from how to do that from observational studies. However, I should warn that these studies require causal knowledge themselves; this holds both for experiments and for instrumental variable analysis. To create an experiment we need to know what the background factors are; we are assuming that there are some causally relevant background factors, and that these are the ones we need to control, but we are also assuming that there are other factors, which are causally irrelevant. When it comes to the instrumental variable approach, we need to make assumptions that the instruments affect the hypothesised cause but does not affect the hypothesised effect. Therefore, we need the causal knowledge to answer this.

Thus to conclude, consider this slogan by philosopher Nancy Cartwright, "no causes in, no causes out". We need causal knowledge in order to produce more causal knowledge. If we fail to put in the right causal scenarios from the beginning, we cannot expect to get the right causal knowledge at the end.

Summary: I argued that causes and correlations are different. I also argued that correlations are evidence for causation, and gave both observational and experimental strategies for how to detect causes.

Part 8 – Engineering Design

8.1 What is special about the Engineering Sciences?

From your perspective as a student in a university, perhaps modern engineering and modern natural sciences do not look all that different. The distinction between universities and technical universities has somewhat eroded. In both, students often finish with similar degrees like a *bachelor of science*. In the industry natural scientists and engineers often enough compete for the same positions. Should we therefore conclude that the natural sciences and the engineering sciences are not relatively distinct from a methodological perspective either? That is, do they share similar goals, kinds of knowledge, methods, and criteria for choosing these methods, so that the scientific methodology that I have taught so far is by and large applicable to engineering too? I do indeed think that a lot of the methodology that I have taught so far is applicable to engineering, however there are some important caveats. The engineering sciences sometimes have different goals, different kinds of knowledge, different methods and different criteria for choosing between these methods. In particular, therefore, engineering science is not just **applied science**, it is not just taking knowledge produced in science and applying that for a particular engineering purpose. A methodology of engineering requires a clear description of these differences, and that is what this lecture is about.

Let me start by sketching in very broad strokes the historical roots of the natural sciences and engineering respectively. What we call natural science today was, up until the nineteenth century, called *natural philosophy*. We find that already with the pre-Socratics in their investigation into the smallest atoms of matter. It is also what Aristotle called his investigations of natural phenomena, as well as Galileo. Isaac Newton titled his famous book *The Mathematical Principles of Natural Philosophy*; Lord Kelvin in the nineteenth century his book *A Treatise On Natural Philosophy.* So science and philosophy until quite recently were not considered separate. Rather, both were seen as kinds of investigation, reasoning and theorizing that is applied to nature and its underlying principles but in similar way it is applied to questions of ethics, political order, human nature and of course also the divine. As a student at Plato's Academy in ancient Athens, in the medieval Muslim or Christian universities, you would have been taught natural philosophy along with these other subjects.

Everywhere the goal of natural philosophy would have been to improve *understanding* of natural phenomena. That is, despite the enormous differences and changes in content, methods and technologies in science the ultimate aim of understanding remained the same. Scientists and natural philosophers acquired knowledge in order to improve their understanding and this increased understanding would drive the quest for more knowledge. So, science and natural philosophy really can be described as an opened-ended acquisition of knowledge for its own sake.

This is very different when we look at engineering practices. Those people taught in the tradition of natural philosophy and working in it where not the ones who built the aqueducts, the castles, or the large cathedrals. The name ’engineer’ enters the English language in about fourteenth century first meaning mainly builder of military equipment and engines. But if we look at the people performing what we now call ‘engineering tasks’ we see that they really have a very different education, they have different goals and there are also different ways of working than the natural philosophers. Take for example Filippo Brunelleschi, the builder of Florence’s cathedral. He trained as a goldsmith and then later worked in a sculptor’s workshop before becoming a master mason and the designer of the cathedral. Or take Thomas Newcomen, the inventor of the machine named after him, an early version of the steam engine. He worked as an ironmonger, providing tools for mines. And James Watt, who refined the Newcomen machine eighty years later, was trained as an instrument maker.

None of these people were trained at the universities in natural philosophy or any of the other philosophical subjects. This is perhaps clearest in the builders of the great medieval cathedrals across Europe. They belonged to the Stonemason guilds across Europe; they were master stonemasons. Their knowledge was accumulated by trial and error and there is no evidence that they received formal training in mathematics or geometry. In fact, it is not known if most of them were even literate. From surviving documents where those master masons made use of mathematics, we see that it is mathematics of a rather primitive kind. It lacks a Euclidian framework and instead exhibits idiosyncrasies of particular masters employing particular techniques that might have been adapted to the particular styles and particular needs of their respective works. So we see that knowledge is sort of organically grown and handed down from individual teacher to individual student.

These craftsmen were trained and organized through the medieval guild system. There was the master mason who developed the plans in accordance with patrons and supervised the actual building process, acting as designer, manager, contractor and bookkeeper, all folded into one. He had the privilege of running his own workshop and he employed so called journeymen, masons who had been trained but who had not acquired a master rank yet. He also trained apprentices and evaluated the skills of journeymen who wanted to become masters. Throughout, this teaching and training was one of “teaching on the job”. As an apprentice, you were following a master mason and you were instructed in person as you were building and working on a building site.

To summarize, here some of the differences that have now come out through considering some of these examples: The sciences aim at understanding. They aim to explain the world and they use abstract reasoning and theory from the start. It is a sort of a standardized theory: in medieval times mainly the Aristotelian works, later increasingly supplemented with mathematical tools. It was transmitted and taught

through books and through lecturing to larger groups at the existing universities.

In contrast, engineering aimed at creating *artefacts*: specific objects, which could be anything from a tool to a cathedral that had particular properties that were satisfying clients, requirements of efficiency, safety, and perhaps also of beauty. So the model of the engineers was not to explain the world, it was to change the world in some way; add something new that would make the world different and maybe a more hospitable place for people to live and work in. That required highly specified and local knowledge that was acquired through trial and error, not transmitted through books and lecturing. Instead, the skills and training that these early masters and engineers required was taught on the job by skilled seniors.

Tacit knowledge: Skills, ideas and experiences that people have but which are not codified and are difficult to transfer to another person by verbalizing it

This shows that what is called **tacit knowledge** was an important component of the engineer. Tacit knowledge comprises of skills, ideas and experiences that people have but which are not codified and are difficult to transfer to another person by verbalizing it. Think of riding a bicycle for example. You might be able and know how to ride a bicycle, but that typically does not mean that you are able to verbally instruct someone else in it, or to write a manual that would then successfully allow other people to learn riding a bicycle. Instead, you typically teach someone riding a bicycle by being present, by showing, by letting the other person try it out, by helping and correcting these attempts.

Another example is how an artesian baker describes a particular feature of his craft. There is much experience lies behind the baker's handling of the dough, yet he has a hard time describing exactly how he does it, he just knows. When he tries to teach this skill he uses an adorable analogy of how you touch someone you love, however I doubt that it will be enough to help the apprentice to really learn. You need to be there, you need to observe the master's time and again you need to copy his moves and then let yourself be corrected while he is there in order to really learn how to handle the dough. Much of craft knowledge is tacit in this sense and early engineering knowledge was too.

Now, modern engineering education began separating itself only about two hundred years ago from these close connections to the crafts by turning tacit, local and trail-and-error knowledge into explicit, general and systematic knowledge. This was done in the first place by opening centralized training institutions for engineers in the end of the eighteenth century, in the seventeenth nineties in France and Germany and in the early eighteen hundreds also at Teknologiska Institutet – the predecessor of KTH – in Stockholm. Finally, these institutions also opened up to women, notably much later than the successors of the natural philosophy traditions, the universities. The first woman in the US admitted to any engineering degree was in 1906 at Cornell (That is almost seventy years later than the first women who received a bachelor degree in the sciences!).

The replacement, modernization and systematization were also achieved, secondly, by successively replacing the workshop with the lecture hall. This was not a step that was immediately taken when these centralized institutions were opened; they initially still relied very much on training in the "shop". It is only in the late eighteen hundreds in the US that there is a push to reduce the shop hours and instead replace them with basic scientific instructions, or science-like systematized instructions. Then, at the turn of the twentieth century, we see that the degrees also become more research oriented. In 1899, the first PhD is awarded in the engineering science in Germany. From that, we then get the concepts of technological science, engineering science, *Ingenieurwissenschaften* and so on.

So to summarize, the goal of the engineering sciences is the *design of artefacts* **(**even though engineers often also engage with the production and maintenance of artefacts). This design requires specialized knowledge bases. In the crafts – the roots of the engineering sciences – these kinds of knowledge are tacit, local and acquired from trial and error. But in modern engineering these knowledge bases are systematically developed, which means that there is a burgeoning engineering theory. There is a division of epistemic labor: not all engineers design, many engineers generate knowledge that is necessary for the time but do not necessarily design themselves. There are systematic variations of possible solutions, systematic testing and different kinds of tests that go beyond the trial and error knowledge. To understand what these knowledge requirements really are and how they can be systematically produced, we now need to look more closely at the main goal: the design of artefacts.

8.2 Technical Artefacts

Artefacts can be many different things, e.g. objects, hammers, a light bulb and genetically modified crop. Engineers, however, also create a lot of processes. Think of assembly lines. An assembly line produces artefacts but is itself an artefact that is designed, can be changed and assessed as being efficient or inefficient. Perhaps this is particularly clear when looking at organic synthesis in chemistry. Think of quinine, a substance used to treat malaria. Quinine occurs in nature as we can extract it from the bark of a tree. But in the 20th century, chemists learned how to synthesize quinine. They created a process that made quinine and it is therefore the process, not the result, which is the artefact.

Systems, at least when they are materially realized, are also artefacts. They consist of assemblies of many artefacts. For example, satellites consist of many objects that themselves are very complicated and designed for the particular purpose of operating within the overall system. This means that it is hard, perhaps, to see what the commonality is between all these different things. It might be instructive to instead ask, what are not artefacts? Or, more specifically for the engineering sciences, what are not technical artefacts? Intuitively, I would say that ideas are

not artefacts, nor are electrons or human hearts. Then there are things that might be artefacts, for instance representations, social institutions or art objects, but they are not technical artefacts.

So, why are these things not artefacts or technical artefacts, respectively? Here are some ideas of what the distinguishing principles might be. First, artefacts are existing material objects. Therefore, an idea is not an artefact (even though it might be a step towards an artefact). I think this is a relatively uncontroversial principle that we can rely on.

Second, many people have argued that artefacts are human made while non-artefacts are in some way "natural". If we think about electrons and human hearts they are, perhaps, seen as "natural" in the sense that they are naturally occurring and naturally produced and would therefore not be artefacts. I am not so sure about this principle. In fact, if we think about the notion of being natural it turns out to be rather complicated. Think for example of so-called incidental tools like Paleolithic hammers. Our ancestors just picked up such a stone and hammered away with it. They used it as a hammer and presumably, that is an artefactual use but the stone itself is natural. It is nothing that has changed about the stone; it is the use that the stone has that suddenly changes.

Then, scientists also make many non-naturally occurring materials. For example, the elements number 95 to 118 are not naturally occurring but scientists have created them. Yet, we do not typically think of those elements as artefacts but they are in some way "non-natural", at least not naturally occurring. Perhaps if we think more broadly of 'natural' we might say that it is something that is not governed by the laws of nature. But this is clearly not true since all the artefacts that engineers generate are governed by the laws of nature; they do not make any exception. Therefore, this notion of natural is very unstable and I recommend that we just do away with it.

Instead, we can use another principle: artefacts are purposeful (they have a function, and the notion of a function will be quite central to this lecture later on), while the identity of non-artefacts does not depend on humans having assigned a purpose to them. For example, electrons exist irrespective of whether someone assigns them some purpose. In fact, we probably cannot tell what the purpose of an electron is; they are just there. Human hearts exist, or have existed, without anybody assigning them a purpose. Interestingly enough biologists now identify organs through their function. Anything that pumps blood through the blood vessels of an organism's circulatory system is a heart. That defines or at least characterizes a heart. But the important difference here is that biologists ascribe this function exposed to components *of existing organisms*. They have not originally created the heart for the purpose of pumping blood, so here the functional ascription is only for understanding. (This does not mean that organisms are never designed nor that they never are artefacts.

In fact, some organisms have been designed for a purpose. Think of breeding farm animals like a milk cow, or genetically modified crops. Those are in fact artefacts.)

Contrast this with the identify e.g. of a hammer. A hammer is not a hammer unless someone describing what she intends to do with it, namely hammering. Therefore, the purpose – the function that is ascribed – becomes one of its necessary characteristics. Artefacts are distinct from non-artefacts by necessarily having a function.

Artefact: Existing material object that has been assigned one or multiple functions.

However, social institutions or art objects might still be seen as purposeful and designed for some function. Are they then artefacts? Yes. Are they technical artefacts? No. Here we are making a further distinction and that is that **technical artefacts** realize their function in virtue of their physical properties.

Technical artefact: an artefact which realizes its function in virtue of its physical properties.

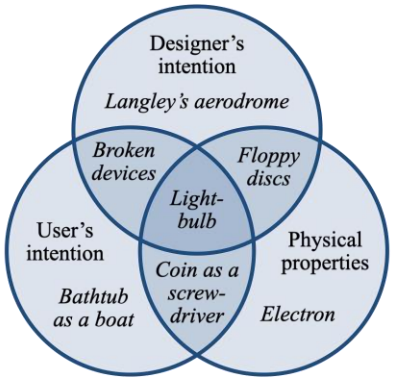
This means that a social institution, such as the National Health Services, is not a technical artefact, since it does not realize its function through physical properties but rather through the common acceptance of a number of rules, for instance the legal code of a country. Such a code can be implemented, real and written down somewhere but how it is written down does not matter. What matters is the common acceptance of such a rule, so it is something else than the physical properties that realizes its function. Similarly with art. The beautiful paintings of Lubaina Himid (that I am using to illustrate many of the lecture slides) have their functions not due to their material or physical properties but rather through the symbolic references that they evoke, which give them meaning. We can distinguish between artefacts generally and technical artefacts, and it is technical artefacts that realize their functions in virtue of their physical properties.

So, technical artefacts are characterized both by their physical properties and by their intended functions. This is what we mean by a technical artefact having a **dual nature**. Edison's patent application of the light bulb nicely illustrates this. On the one side, he describes the intended functions like giving light, incandescence, offering resistance, providing flat surfaces and so on. On the other side, he describes the physical properties like the Corbin wire, platina wires, the vacuum bulb, and the glass receiver.

Dual nature: technical artefacts are characterized both by their physical properties and by their intended functions

Both of these together describe his new invention. Importantly, he needs both to characterize his technical artefact. He could not just use the functional properties, nor could he simply use the physical properties to describe what he had invented. Nor can he reduce one to the other. It is not possible to logically infer from the physical properties what functional properties such a device has or vice versa. There are of course constraints. For something to be a functional device, it requires certain physical properties, or at least some arrangement of physical properties, and if it has certain physical properties then it cannot have certain functions. But,

these two realms cannot be reduced to each other and we cannot therefore rely just on one of these realms in order to describe technical artefacts.



Intended function/application/physical properties (thanks to Jonathan Fridstrand)

Let us have a look at some difficult cases. Here, I am thinking about artefacts that malfunction, like a broken screwdriver. Or, what should we say about inventions which fail from the start because the inventor lacked expertise or competence? What about incidental tools? For example, something you just picked up and you found that it is useful for something you currently need. An example would be a situation where you do not have a screwdriver but you have a coin in your pocket, which you use to move the screw. How about misuse? For example, a comical misuse would be someone using a bathtub as a boat. Lastly, what about artefacts which used to be really important but are now obsolete like the floppy disc?

In order to make such examples clearer we can use the dual nature of artefacts. But we need to introduce a further differentiation, namely that between designers' and users' intentions. That gives us three overlapping but at least partly separate domains, namely the designer's intended function, the users' application and the physical properties. These domains overlap in a Venn diagram and we can fill in each of the gaps.

First, where they all overlap: a designer intending a particular function that the user then makes use of, and the physical properties supporting such functions. This is the standard case of the artefact, for example the light bulb. Second, we can look at the outer circles where we find objects that are only placed in one of each domain. The first examples are those objects that only have physical properties and lack any function. These are the non-artefacts, like the electrons or the human hearts.

Then, there are the failed inventions, and here I am thinking of Langley's aerodrome that you have already encountered in the end of the modeling lecture. This was an idea of what sort of functions a certain physical object could have, in particular: a vehicle that is carried by the air. But it failed: it fell straight into the Potomac River. Then we have the misuses, e.g. the bathtub used as a boat. These cases are interesting because the bathtub, of course, had a designer's intended function but a very different one, and the user employs it in some very different way that is not supported by physical properties. In these cases, I would argue that we are not even talking about artefacts - even though we will probably need to differentiate them a bit further later on.

But there are also all the partial overlaps. These are interesting because they are artefacts but they might fail in some normative sense. For

example, the incidental tools: the coin that we used to screw a screw is an artefact but it is an artefact that we cannot really say malfunctions, or is misused in some way. I will come back to this later. Then there are the malfunctioning devices: things that the designer intended for a particular use and the user wants to use it for this particular use but have somewhere along the way lost the physical properties that would support this intended function. We still have expectations and normative claims towards such malfunctioning devices and so they are artefacts but they are somewhat deficient. Lastly, there are the obsolete devices that still have the physical properties that support the designer's intended function but the users are not using it anymore.

To summarize, artefacts are material objects that have a function intended by its designers and its users. They realize their function in virtue of the physical properties, and such technical artefacts may malfunction, they might have incidental uses and they might become obsolete.

8.3 Defining Technical Functions

Function: the purpose that someone intends for an artefact

What are **functions**? According to the dual account, artefacts are characterized both by physical properties and functional properties. As I argued, functions cannot be reduced to physical properties, so they cannot simply be the capacity established by the physical properties. Instead, functions are established by peoples' intentions. In other words, the purposes that someone intends for an artefact; which functions an artefact ought to have. Here you can see how we are transitioning from functions to the normative quality of an artefact. This is what Herbert Simon (*The sciences of the artificial*, 1969) means when he says that "the scientist is concerned with how things are but the engineer with how things ought to be", considering a normative quality that is absent in science. This is what I want to discuss now.

Normative: How things should be, regardless of how they are.

The **normative** quality of artefacts is clearly indicated by questions like, "What does it mean that an artefact malfunctions?" In other words, what does it mean that an artefact does not function like it should? What does it mean that an invention fails? It does not do what it is intended to do. Or, what does it mean that artefacts are misused? They are used in ways as they ought not to. Natural scientists, who only focus on physical properties of objects or phenomena, do not encounter such normative claims. So, the conclusion might be that these normative qualities must emerge from the functional properties specific to engineering and the artefacts engineers engage with. How can that be?

If it is peoples' intentions for the objects that establish these functions, and thus their normative quality, then there are two questions that arise in particular. The first is, "Whose intentions count?" Maybe some peoples' intentions count more than others. Then, once we have identified whose intentions, the second question is, "Do any intentions establish a function and thus a normative claim; irrespective of how unfounded they are in the physical properties?" I now want to discuss these two questions, so let us start with the first one.

As I have already discussed in the previous sections, artefacts and their functions can be established either by a designer or a user. In some difficult cases these things might come apart, a designer might intend something while a user might use it in a different way. Or, there might not exist a designer at all who intended anything and the users just pick up an object and use it in some way. So we can distinguish two ways of how functions can be established and that will give us a key to the questions of normativity. The first one is that functions can be established by **ascriptions**. That is, a user who believes that an existing object has certain capacities and therefore ascribes a function to this object. Think of the coin that I might find useful in order to turn the screw. The coin was designed for some purpose but not for the purpose of turning a screw, and yet I as the incidental user in a particular context might find that to be useful. So I am ascribing the function of screw turning to the coin.

Ascribed function: a function given to an existing object by a user who believes that this object has certain capacities that can fulfill a certain purpose

This is an epistemic question: my belief that the coin can perform the function of turning the screw or not can be correct or incorrect. My belief can be incorrect in the sense that I believe that the coin can turn the screw but in fact, it cannot – e.g. the screw slot is not wide enough. However, I would not say that this situation establishes a malfunction. It is not that the coin malfunctions; it is simply that my ascription was wrong from the start. Nor is there a sense of artefact misuse. It is not as if the coin's use as a screwdriver is something that is prohibited from the designers of the coin; it is just that it was not part of the original use plan. So we do not get any kind of normative content form this kind of ascription as establishing a function, and therefore we can think of an ascription as a generally weak form of how a function is established.

Contrast that with the **function assignment** that a designer might perform. A designer intends to describe or create an artefact with a certain function, and based on her expertise then describes or creates the artefact. By doing so, she performs an act that might establish a new instance of a **functional kind**. Before 1879, many people had an idea of what electrical light was and should do. Then Edison created a concrete instance of this functional kind: the carbon-filament, vacuum-based, high-resistance light bulb. This artefact as he described it was supposed to fit the functional kind of electrical light, something that produces visible light from electrical current. It is not the only way in which it could be done, but it is one way of how it could be done (or so Edison claimed). Similarly, we might say that a new design for an airplane is an instance of the functional kind that flies by gaining support from the air. By establishing an instance of a functional kind from these functional assignments, we can then derive a stronger normative claim. Namely, because we identified an object as an instance of a functional kind we generate the normative claim that this object ought to satisfy the function of this kind.

Function assignment: a function given to an object by a designer by creating or describing an object with certain capacities that can fulfill a certain purpose

Functional kind: A categorization of artefacts and their descriptions in terms of the functions they ought to satisfy

With these concepts, we can now explain the notions of malfunction and misuse. A *malfunctioning artefact* is an artefact that has an assignment such that it ought to satisfy certain functions of a given functional kind, but which has now have lost the physical properties that did support its function. It no longer has the physical properties needed in order to satisfy

the function that it was assigned, that it ought to satisfy. A *misuse of an artefact* is one where a user employs an artefact in a way that does not realize the function of its functional kind (i.e. the functions assigned to it).

Now, we have two answers to the question of whose intentions that matter through two ways of how a function is established. One is through the ascribed function that has a weaker normative claim and the other is through function assignment that has a stronger normative claim. But we nevertheless need to qualify this stronger normative claim because not all of a designers function assignments will work. This means that a designer might fail to assign a function to an artefact if the physical structure in a fundamental way is not capable of supporting the function. Only those assignments that are backed by substantially correct expertise (design knowledge) establish a function.[3]

Thus, the normative power of functions really comes from the designer's assignment, not from the user's ascription. We can easily see this in the standard case of the light bulb. It satisfies the designer's assignment, the user's ascription and the physical properties. But we can also see that the normative claim is established even for cases of obsolete artefacts, as well as for malfunctioning artefacts. In each case, it is the designer's assignment that drives the normative claim that is inherent in these two kinds of artefacts.

To conclude, the function of an artefact is established by either designer assignment or user assignment, and it is competent designer assignment that establishes the normative claim. This, in turn, allows us to characterize malfunction and incidental use, while the functional ascriptions by user do not establish a similar normative claim.

8.4 Knowledge for Design

All artefacts are material objects, and therefore they all operate through cause-effect relations that are governed by laws of nature. And scientific knowledge is of course helpful in order to understand such artefacts. Why does an airplane fly? What produces nuclear power? Why does aspirin reduce headaches? Answers to such questions typically would refer to scientific principles. Yet, scientific knowledge alone cannot support design or the design process, as I now will argue.

---

[3] Langley's aerodrome is an interesting example here because it did fail. It was an invention, or at least an attempted invention, of an airplane but the artefact that was supposed to satisfy the functional kind of an airplane actually never flew. However, even though Langley was wildly ridiculed for his failed invention at the time, it was based on a lot of expertise which is shown by the fact that the smaller scale model of the airplane did function according to the functional ascription. It was only the scaling up that did not work. Even today the Smithsonian that has collected Langley's surviving exemplars of the aerodrome thinks that it is an artefact even though it failed to function as intended, so we have some leeway here. We might say that Langley showed substantial expertise in developing his invention, thus performing a functional ascription that at least partly was established as an instance of a functional kind even though it did not fully satisfy it.

8.4.1 The design process

So, let us first have a look at what the design process is that requires this different, non-scientific, kind of knowledge.

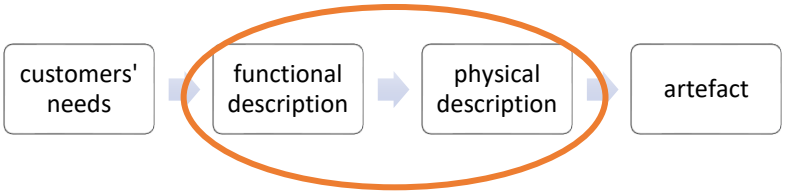


Figure 1: the design process

The **design** process consists of a number of steps, going from the customers' needs, through the functional description, to a physical description, to the production of an artefact. We can say, more succinctly, that design generates an artefact by bridging the functional and physical descriptions. It generates the dual nature by producing or developing *physical description* that corresponds *to functional description*. That might be the core of what said design really is about. It is a little bit purist, because typically, we would think of design as ending in the production of an artefact, and for that, the definition needs to be widened.

Design: A process that develops a physical description corresponding to a functional description.

So far, the core only concerns the development of a physical description. Instead, we might have a slightly wider, **creation-centered** definition, where design is a process that creates a physical object that robustly satisfies a functional description. Then we might say that it's also very unclear where these functions come from, so the designer also needs to take into account the **client's** needs, or, widely, the analysis of the functional requirements. Including that would mean that the design process is defined as a process that creates a physical object that robustly satisfies the client's needs. But what is crucial here is that at each of these steps (represented by the three arrows in figure 1), the designer needs to refer to particular knowledge, and this knowledge often needs to be generated; it isn't just there; it is part of the design process; it is part of the helping and supporting of the design process that such knowledge is generated. Let me now go into three of these knowledge areas and give some examples and subcategories.

Creation-centered: A design process creates a physical object that robustly satisfies a functional description.

Client-oriented: A design process creates a physical object that robustly satisfies the client's needs.

8.4.2 What functions should the device satisfy?

The first kind of knowledge that is required under the wider definition of the design process is to know: *what functions should the device satisfy?* That is typically not given. You might think that the client gives you a full list of design-required functions, but in fact, the client often doesn't do anything like this. Here's a navy official – obviously involved in talking to engineers on a regular basis – who complaied in 1941 that the

navy was not properly doing its job of instructing engineers what artefacts they should design:

> "At present we simply specify that the airplane shall be perfect in all respects and leave it up to the contractor to guess what we really want in terms of degree of stability, controllability, manoeuvrability, control forces etc. He does the best he can and then starts building new tails, ailerons, etc. until we say we are satisfied" (Capt. RS Hatcher, Navy Bureau of Aeronautics, April 1941).

So this is a bit of a blind flight from the engineer's perspective: you are not given a full functional specification, but instead you have to develop that yourself. Of course, in the early days of the development of the aircraft, there was not just that information available somewhere; it had to be generated. You had to collaborate closely with the big clients, like the navy or the air force or other people, who wanted to put airplanes to commercial use. But you also had to collaborate closely with the pilots – the people who were actually using these devices – in order to see what it even meant that your airplane has a degree of stability or manoeuvrability.

So you need to develop the functional concepts to start with; understand what those dimensions of functionality were; and then you need to learn what, for each of these dimensions, your clients and your users, wanted. Those were important parts of information that the designing engineer needed to require, and there is no real corresponding knowledge that a scientist would at all have, or even aim to have, because she would not work in terms of functional kinds at all.

8.4.3 Which physical structures might satisfy the demanded functions?

Going to the next transition – the transition from functional descriptions to physical descriptions – the question now is: *which physical structures might satisfy the demanded functions?* And here, of course, science helps a lot; but if you look more closely, you see that, yes, engineers often rely on theory, but the kind of theory that they use often is not scientific theory. So, on the highest level of abstraction, the designer trying to answer this transition question from functions to physical descriptions will refer to **operational principles**.

Operational principle: The underlying, basic concept of why a particular physical structure might deliver a certain functionality.

For example, what is an airfoil and why does it function as it's supposed to do as part of an aircraft? That is something that has been established, but it has been established not by scientists (even though scientists could then explain *why* this would be the case), but by engineers, who crafted artefacts that satisfied functional kind of an airfoil. That established an operational principle - which is a particular kind of engineering knowledge.

On a slightly more specific basis, the engineer interested in building an aircraft might then refer to other specific design or engineering knowledge in the form of *normal configurations*. Even in the 1940s, there were lots

of different ways how airplanes had been built until then. Examples are aft tail vs. frontwing (canard) designs; and monoplane vs. biplane designs. These are all examples of the general shape and arrangement that are commonly agreed upon as to best realise a particular operational principle. Again, a kind of knowledge that engineers employ on a day-to-day basis, but that scientists would not typically be engaged with.

More specifically, engineers also engage with *device-restricted simplifying theory*. For example, classic beam theory is a simplification of non-linear beam theory, which is itself a simplification of continuum mechanics, and it is applicable to some kinds of devices, namely those where deflection and depth of beam is small compared with span. That is extremely useful for an engineer who is building within those confines, while for a scientist, aiming for more general principles, such a device-specific simplifying theory is not very useful.

Engineers even engage in more *surface-level theories*, that don't even give a nececessarily correct account even for restricted domains. Think of theories of eddy viscosity, a concept used in modelling turbulence; they ignore a lot of what is really going on in fluids, but it might still work for the purpose of calculating large-scale motions that ignore the small-scale vortices. Now, we are really deviating from theory and the physicist might say that (a) this is not rigorous and (b) it might actually be faulty. Nevertheless, engineers use it, because it is enough for some of their pruposes. Sometimes, you simply go about in a way that we shouldn't really call theory. It's just some kind of simplifying quantitative assumptions. For example, that the loads between a number of rivets are equally distributed. There's no basis for that, it is just a way to simplify calculations at some point that might be useful.

So we have a lot of theory that is the engineer's own, and that scientists would not engage in or participate in producing. That is the basis for transitioning from functions to physical descriptions. But engineers also have a lot of *data* that they produce. There are catalogues and handbooks full of measurements of particular device components; for example, the catalogue of airfoils produced by NACA in the 1920s, where the performance of each airfoil has been tested in a wind tunnel. So now, as a designer with a particular set of functional requirements in mind, you can go to this catalogue and check what would be the most useful physical description for your purposes.

8.4.4 How to produce these physical structures?

Now to the final transition. The one from physical description to artefact: the production knowledge. There is a lot of knowledge about *materials*, about *labor force*, about *expertise availability*, that you need to have in order to determine how to produce something that you have now physically described. There are *efficiency considerations* and *scaling considerations* that you need to take into account. There are also *practical considerations*, for instance, how much clearance do you need to calculate

for an assembly worker to actually reach inside of a device in order to install something particular. All of these things, again, are particular engineering knowledge required for the design process that will not be available from a scientist.

We can now distinguish between science- and design-knowledge, contrasting the two on a number of dimensions. The engineer is interested in *investigating functional demands*. The *scientist does not have a concept of technical function* and is not interested in it. The engineer aims to work with *very concrete functional objects*. She wants to make the functional description as complete as possible (we will come back to that later), while the scientist, beyond the general goal of understanding, and perhaps the sub goal of explanation (or kinds of explanation) and description, will *not have very specific concrete objectives*.

Engineering knowledge	**Science knowledge**
Investigate functional demands	No concept of technical function
Work with concrete functional objectives	Few rigidly specified goals beyond understanding
Organize theory and data around functional kinds	Organize theory and data around natural kinds
Allows for device- or context-specific knowledge, where convenient	Always aim for generality
If the devices to which they apply cease to be useful, the knowledge will no longer be of interest	Once acquired knowledge always of interest

The engineer will organize theory and data around *functional kinds* – we saw this in the various examples of engineering and design knowledge – while in science, you will organize theory and data around *natural kinds*. So the kind of categorization is very different in the two fields. The engineer will allow for a lot of local knowledge – *device-specific knowledge, context-specific knowledge* – wherever it is convenient and gives sufficiently good results. The scientist aims at *general principles* as much as possible. The engineer, finally, will be interested in the knowledge that has been produced in the past only to the extent that the devices, for the design of which this knowledge is important, are still relevant. *If those devices become obsolete, the knowledge will no longer be of interest*, so it can be discarded. There is a lot of past engineering knowledge that is of no interest anymore. That does not hold for science. In science, anything that truly is knowledge – that is true, justified belief about the natural world – will remain of interest in order to understand the natural world, and *there is no sense in which that knowledge becomes obsolete*.

> "The scientific knowledge of a machine as an object tells us nothing about it as a machine." (Michael Polanyi, 1958)

Thus, we see a real separation between engineering- and science-knowledge and these many instances. As Michael Polanyi said succinctly in the 1960s, the scientific knowledge of a machine is a very different

knowledge than engineering knowledge, and it "tells us nothing about it as a machine" itself. Rather, scientific knowledge only treats it as an object amongst other natural objects and it completely neglects that other side of artefacts, which is its functional side. That functional side is the domain of engineering knowledge.

8.5 Design Methodology

I have argued that design requires its own, special kind of knowledge. There are many methods that could generate and apply such knowledge. What justifies choosing some of these methods over others? That is the topic of the methodology of design. Let us go back to the simple model of the design sequence that I described earlier. That gives us the steps through which the designer proceeds.

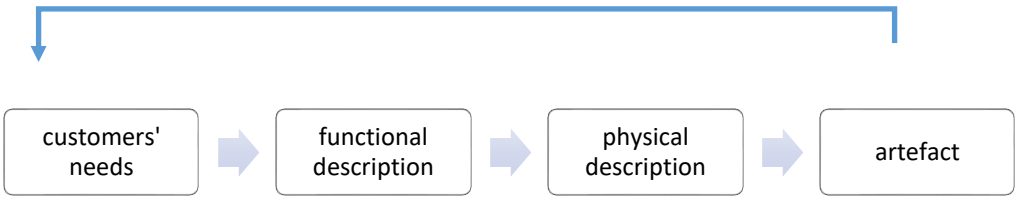


Figure 1: the design process and method choices

Different method choices must be made at each of these steps, and here, although without question in an incomplete way, I will discuss four.

The first is – what kind of functional description should be given, and what is an appropriate functional description? The second is – how to generate and weed out possible physical descriptions that satisfy the functional requirements? The third is – how to select, from these possible descriptions, the most promising one? The fourth, finally, is – how to validate the design that has now been described, or even produced a material artefact?

1. What functional descriptions?
2. How to generate & weed out possible physical descriptions that satisfy functional requirements?
3. How to select the most promising possibilities?
4. How to validate design?

8.5.1. Appropriate functional requirements?

Start with the question of what is an appropriate functional requirement. Here, one can at least impose four conditions. The first one perhaps is most obvious, saying that it should be a *functional description*, that is, it should describe what it is *for*, but it should abstain from already giving restrictions on its physical implementation – so it should not describe what it is made of. That is a negative requirement. But then, the positive requirements are that it should be *complete*: does the description describe all desirable effects of the artefact, or are some important ones left out. Of course, we do not know. In some way, this is an open-ended quest, say, investigating all the users and the clients who want to make commercial use of it etc. but perhaps you're missing out on something, and you will not find conclusive evidence that you have now hit the complete effect-list, so this is a very demanding requirement. Another one is that the functional description must take care of the *use- context*. The functional description depends on the users' behavior, and the way in which users interact with the device. Finally, a good functional description specifies *threshold of failure and degrees of success*, so that later on, in the validation of design, you can refer to such criteria in a reasonably precise way.

Implementation independence: avoid specifications of physical implementation

Completeness: describe all desired effects of the artefact

Use-context: make functional description dependent on users' behavior

Quantitative criteria: specify threshold of failure and degrees of success

8.5.2. Generate a set of physical descriptions

How to generate a set of physical descriptions? The standard practice here is *functional decomposition*. We start out with a high-level description of a function, and then we analyze that function into a set of sub-component functions that together constitute this higher-level function. Maybe on this level, we begin to find (at least for some of these sub-functions) that there are standard solutions, solutions that have been implemented before. Here we might refer to the operational principles or the normal configurations, so, when building an aircraft, perhaps the landing gear, or the way in which the landing gear is lowered upon descent, is something that has already been provided with a normal configuration satisfactory for your particular requirements. Then, you do not need to do more than functional decomposition that leads you to this standard solution. This might need to be reiterated on a number of levels, but even then, you will probably not be able to find standard solutions for each of these sub- or sub-sub-functions.

Instead, there might be some sub-functions that do not have a standard solution, and here, you need to give a physical description without a clear knowledge base. In this case, you might call it blind. It is a blind

- Functional decomposition
- Blind variation
- Weeding out through modeling or partial experimental test

variation: you need to try something new. How do you do that? Well, blind here does not mean that it is just random or unconstrained. Even if you lack the knowledge base for how to physically implement it, you might have a lot of knowledge, both theoretical and empirical, about what *does not* work, and that will constrain within which limits you can perform this variation. Furthermore, you might engage in analogical reasoning, using other artefacts and their operation in some way to help you specify the details of the physical descriptions for the sub-functions that you have to vary blindly. Finally, it is important to take the *variation* part seriously to, namely that you don't just end up with one specification within the reasonable constraints, but that you try to cover a number by, for example doing what in some cases might be called *sweeping the reasonable parameter space*, or something of that sort.

Of course, you then also need to engage in weeding out. We do not want hundreds or thousands of alternative possibilities. Here, you might engage with either theoretical modeling, using, for example, a simulation of parts or the full artefact in order to compare the relative advantages of such variations. You might then go on to even materially model part of the artefact, as, for example, the Wright brothers did when they tested the propeller shape of their aircraft in a wind tunnel.

8.5.3. Decide between possibilities

Once you have assembled a number of possible alternatives and have begun to test, either theoretically or materially, the relative advantages and disadvantages of these alternatives, you need to have some decision procedure. Here, most engineers tend to think in terms of *optimization*; choose the one that performs optimally with respect to the relevant functional requirement. But optimization, properly understood, requires a separation of *search* and *evaluation*. So you first perform an exhaustive search. You generate a large number of physical possibilities, you evaluate each of these alternatives separately, and then you determine which of these is the best one.

That is very systematic and exhaustive. However, it is also very effortful. Instead, what people often do in practice is what Herbert A. Simon in the 1950s called **satisficing**. This is a combination of the words 'satisfy' and 'suffice'. The idea here is that you do not separate search and evaluation; instead, you combine them. You define a certain minimal threshold. Think of when I talked about the functional description and the minimal criteria: we needed some form of threshold for failure. That might be the

Optimization

- Exhaustive search
- Complete evaluation

Satisficing

- Determine minimal functional threshold
- Search until threshold is met

Satisficing: A process where search and evaluation is combined, so that options that fulfill a certain theashold are chosen.

threshold that you are referring to here. Now you might start search, and as soon as you come up with one possibility, you evaluate this possibility with respect to the threshold. If it is above the threshold, then end search and take that option. Alternatively, if it is below the threshold, then continue. This might be much closer to practice, and there are certainly many context in which such a satisficing decision-procedure is more effective than full optimization.

8.5.4. Design validation

Last, once you are beyond the decision and you have produced at least a proof of concept or a prototype, you might perform testing. These might be, as I said, forms of prototype testing, or it might be a full production with direct trial as well as usability testing. In those cases, you will use empirical methods – like material modeling, experimenting, and observational studies – known from the sciences. On the lower level, you will use criteria like, for example, constructive validity for measurements that you perform in such empirical practices. The criteria are the same as in science, but the overall goal is different. They don't to show that something is true. Rather, they aim to show, or to investigate, whether an object satisfies the relevant functional requirements.

Design validation

- Proof of concept
- Prototype
- Direct trial
- Usability testing

Design validation: Testing aimed at ensuring that a product or system fulfils the specified functional requirements under specified operating conditions.

8.6 Summary

To summarize, design, at its core, is the translation of a functional into a satisfactory physical description. Of course, engineers often draw on scientific knowledge, but it is neither enough nor of the right kind to perform design by itself. Instead, other kinds of knowledge are necessary: functional demands, engineering theory, engineering data, practical considerations, and so on. Therefore, engineering design requires its own knowledge, and it is just wrong to categorize it simply as applied science.

Even the simple sequential model of design identifies various points of method choice. Overall, there is little agreement on how to justify these method choices in general. I think that the methodology of design and its discussion is not something that philosophers have paid a lot of attention. Nevertheless, I have been able to identify a number of criteria that might be relevant. I have also stressed again that these method choices, as well as the criteria on which they are made, are different for engineering design and for science.

Part 9 – Qualitative methods

Reading instructions. Some sections in this chapter are only intended for certain student groups.

- Section 9.1.1 is an introductory section for social science students only.
- Section 9.1.2 is an introductory section for all other students.
- Section 9.6 is only intended for social science students.

9.1 Introductory sections

9.1.1 Naturalism in the Social Sciences

When discussing social science methodology, many stress the differences between the social and the natural sciences. Indeed, there are substantial differences. In the first place, the social sciences unlike natural sciences focus on human behavior, the origins of human behavior, human interaction and the social institutions and organizations that emerge from these behaviors. But social science itself is also a human activity, so it is humans investigating humans. Therefore, there is not the kind of difference between the investigator and the investigated subject in the same way as in the natural sciences where the investigator and the investigated object typically are of entirely different kinds. Instead, social sciences themselves are a form of human interaction. One interviews, questions, debates, interprets.

Qualitative data: To be added.

In many cases, social scientists produce a different kind of data, namely **qualitative data**. They produce sound recording, video recording, diary entries, experience reports, open-interview answers, questionnaire answers, photographs and so on. Now, these really play an important role in the natural sciences where we are used to quantitative data by and large. Social scientists also focus on a small number of cases in many instances. That is, instead of sampling representatively from a population they pick out a few cases and study those cases in depth, thus gaining depth but loosing representativeness. Finally, social scientists very often interpret, that is they try to understand behavior by assigning beliefs, desires and intentions.

More specifically we can look at the various social-science methods that social scientists use that are often quite different from the methods used in the natural sciences. For example, social scientist propose questions to subjects in interviews, questionnaires and by asking a focus group to debate a particular subject or a particular issue or question. On the other hand, we also investigate humans in their context as they behave, like someone sitting on a chair observing someone making use of his/her kitchen (as in the Swedish movie Kitchen stories). Or sometimes (actually very often), social scientists also construct laboratory conditions under which they observe human behavior, for example in a usability experiment. Often, we are making these behavioral observations not just to record behavior but also in order to learn something about the

underlying motives, emotions and feelings of people, and then we are helped by asking the subjects to think out aloud. We let them say what they are thinking – what is going on in their minds – as they are behaving in a particular way.

These methods are particular to the social sciences. We do not interview electrons or send a questionnaire to non-human animals. We are also making inferences that are different from those inferences in the natural sciences from the data produced. For example, we infer to the beliefs, desires and intentions of our subjects. Or we conclude about personality or character traits, like the Big five (openness, conscientiousness, extraversion, agreeableness, neuroticism) that are famous in psychology for example.

Therefore, I think there can be no debate that the social sciences indeed use methods different from those used in the natural sciences. The questions now is, given that social sciences make use of different methods, do social sciences therefore also need a different methodology? Remember, I have distinguished between methods as strategies and processes for conducting research, and methodology on the other hand as reasons and criteria for choosing between these methods. Here we can distinguish two views. The **anti-naturalist view** of social science says that because the social sciences' methods are different – the qualitative in the social science versus the quantitative in the natural science, interpretation in the social science versus causal regularities in the natural science, the nomological generalizing versus the ideographic particularizing in the social sciences – therefore the social sciences' methodology must be different too. Thus, the anti-naturalist position concludes that the social sciences cannot and should not follow the methodology of the natural sciences. Against that, we can contrast the **naturalist view** of social sciences. The naturalist view admits that the social sciences do indeed use different methods in many of the cases but it asks, "Why should not social sciences' methods be chosen according to the same reasons – the same quality standards really – as other scientific methods?" After all, social scientists pursue the same ultimate goals as natural scientists. They aim to explain and predict the phenomena of interest and they aim for reliability of methods and the validity of results. If these goals are the same then perhaps the methods that they chose in order to reach these goals should be judged by the same standards even though the particular technicalities of these methods differ in one way or another. Therefore, the naturalist view of the social sciences claims that the socials sciences can and should follow the same methodology as the natural sciences despite their differences in methods.

The anti-naturalist view (of social science): Since the methods of social and natural science differ in so many ways, the social sciences cannot and should not follow the methodology of natural sciences.

The naturalist view (of social science): Since social science shares the goals of natural science, the social sciences can and should follow the methodology of natural sciences.

Of course, at this point, I have only given you the contrast and I have not given you the argument. This is one of the main overall goals in this lecture. I want to give you arguments for the naturalist view in the social sciences. To show that, how the most important of these special methods are nevertheless subject to the same methodological standards and evaluation criteria as the natural sciences' methods are, as I have discussed in previous lectures.

9.1.2 Studying Human Behavior

As I have argued in the previous lectures, design often involves the studying of human behavior. For example, the designer investigates what functions a device should satisfy by examining the potential clients' demands, desires as well as the early users' experiences and requirements. An example is that of test pilots. Then, towards the end of the design process, the designer needs to validate the proposed design. That is, to check whether the design product fulfils the functional requirements by examining the users and their experiences, for example in a usability experiment.

All of this requires investigating human behavior. In some cases, one can use natural science's methods for the purpose of studying human behavior. For example, when one investigates the purchasing decisions that customers or potential customers make in order to find out how good a new product that one has designed and brought to the market is. This can be done by using standard statistical methods just as if we were studying a weather phenomenon; nothing special here. However, in many other cases, we need to investigate not just human behavior but human feelings, desires, plans, beliefs and their intentions. For this, we need to apply particular methods that are quite different from those of the natural sciences.

We can imagine an investigator sitting on an elevated chair observing a purported kitchen user (the idea is taking from a movie) in order to find out, not just how the person behaves, but also what kind of atmosphere the kitchen provides, what the person feels and whether the person is satisfied with the way this particular kitchen is designed. In other cases, researchers provide laboratory environments in which they observe people's behavior, for example in a usability experiment. Often, this is supplemented by asking people to not just perform a task but also comment on the task as these comments can reveal what their feelings and thoughts are.

Such comments can be very revealing, as we can see how an agent tries to actually make use of a particular device but fails. The behavior itself only shows that the agent stops, that the agent does not know how to proceed. But by letting the agent reveal what his/her thoughts are, why he/she is surprised or frustrated, we learn a lot about what is actually working and what kind of thought process that is triggered by the device. We can also investigate this without making people behave. We can let people behave in some environment in which we send them a questionnaire for example. In this way, we let people answer in the comfort of their home. Alternatively, we can call these people or visit them personally in order to perform an interview, or we put them in a group of people and let them discuss a particular view and see how they react to each other.

All of these methods get us particular observations and data that is different from the data we that we get from the natural sciences. Furthermore, from this data we make then inferences to particular

conclusions that are different from those in the natural sciences. For example, in a usability experiment we infer from people's behavior to their beliefs and desires. Alternatively, we might make use of an in-depth-interview study in order to learn something about people's personal character traits. Think of here of the Big five (openness, conscientiousness, extraversion, agreeableness, neuroticism). These are examples of what we infer to, what we want to know and what we might hope to conclude to from these methods.

The methodological question now, however, is how to assess these methods and how to justify choosing between them for particular purposes. Some people claim that these methods are so fundamentally different from natural sciences' methods – that they are qualitative rather than quantitative, that they perform interpretation rather than identify causal regularities, and that they focus on the specific rather than trying to reveal general regularities – that the social sciences' methodology therefore also must be different. I will argue against such a claim. I will show that, although these methods indeed are different from those methods used in the natural sciences, the reasons and quality criteria for choosing between them and for assessing them are in fact the same. That is because we pursue the same goals with these methods as natural scientists do in their respective domains. Social scientists want to explain and predict their phenomena, they want to use reliable methods for these purposes, and they want to achieve valid results just like natural scientists do. Therefore, in this lecture I will argue that these methods in fact can be subjected to a similar methodological assessment as the natural sciences' methods that we have discussed before have been.

9.2 Interpretation as a Scientific Method

Interpretation is a central method for the social sciences. It is based on a general human capacity that we all automatically and pre-theoretically seem to have. If you look at paintings of human faces, you do not just see the paintings of human faces. You see the painting of humans with certain emotional and mental properties, you see someone who feels, who intends or desires something. We cannot really extract ourselves from these observations since our attributions of these mental states to these faces, to the depicted persons, are rather automatic. How we perform this capacity is not so clear; there are two different camps that can be broadly distinguished. The **simulation theorists** argue that we reuse our own cognitive mechanisms – our way of believing, desiring and intending – to simulate other people's mental states. So we are occasionally switching from our own desiring to simulate how other people desire or believe in situations when we put ourselves into their shoes. Against that, the so-called **theory theorists** argue that we construct a theory of mind, just like scientists construct a scientific theory. We collect evidence, formulate explanatory hypotheses and revise these hypotheses in the light of further evidence, now about other people's minds.

Simulation theory of interpretation: We reuse our own cognitive mechanisms to simulate other people's mental states

Theory theory of interpretation: We construct a theory of mind just like a scientific theory, using collecting evidence and revising hypotheses.

This so-called folk psychology has been used as the basis for a lot of social-science research. It is not the same, as we will argue later, but we

can see origins in a lot of social-science research in this common sense or folk psychology. Think about interpretations in science, for example in a usability experiment where a subject has been asked to play a computer game and a researcher is recording the subject's behavior and maybe also the subject's utterances, as well as the results from her interactions with the device. In the first place, what the observer sees is the appearance of an instruction and the resulting choice of the subject's behavior. But instead of merely trying to connect those two, what the researcher does now is to attribute a number of mental states to the subject. So, the subject gets an instruction and then the researcher interprets the subject as understanding this instruction in a certain way, in her way of understanding what this instruction means. Then, the subject is interpreted as having a number of beliefs – for example, what her choices are given the instruction, or what effects her choices have given the environment that this particular gain offers her. Then, furthermore, the researcher now also attributes desires to the subject, for example, that she wants to win the game, or perhaps that she wants to annoy the researcher and sabotage the whole exercise. But it is only through this chain – from instruction through understanding, through believes, through desires to the choice – that we then interpret the subject's behavior and come to understand it.

This understanding, then, is achieved through a so-called **belief-desire explanation**. We see subjects, or we record a subject's choice and we explain that choice by pointing to her relevant believes and relevant desires. So because the subject desired or preferred state X over state Y, and because she believed that action A brought about X but not Y while action B brought about result Y but not X, therefore she choose action A over action B. So interpretation is something else altogether than merely correlating input and behavioral output. Rather, we assume that these inputs, the instructions in this case, are mentally processed by the subject leading to behavior that is meaningful in the light of the subject's subjective understanding, believes and desires. The social sciences' interpretation involves a justified assignment of such understanding, believes and desires in order to make the observed behavior meaningful.

Belief-desire explanation: An agent's action *A* (instead of *B*) is explained by her desiring some result *X* (and not *Y*), and her believing that *A* will bring about *X* (and that *B* will bring about *Y*).

Such interpretation matters, not only in the social sciences but also in other non-scientific-social contexts. Think for example about the claim:

The accused murdered the victim by running him over with his car.

What makes such a claim true or false? I want you to think about this for a moment and try to answer it before reading further.

This kind of claims is a problem that a criminal court faces all the time. Whether it is a judge system or a jury system, the accused is being interpreted. The accused is not simply being held for his behavior, but rather for the intentions and desires that he had and that led to this behavior. So to take the example, if the accused is interpreted as having the desire to kill and having the belief that his actions would kill the victim, then he will be (or is likely to be) indicted for murder. However, if he is interpreted as lacking this belief and desire then he might be

indicted for manslaughter or for negligent homicide, and that makes a huge difference. In Sweden for example, if you are indicted for murder, you will go for prison for at least ten years, but if you are indicted for negligent homicide, you will be going for prison for maximally two years. And this depends entirely on how your behavior is interpreted, not on what your behavior is. You ran over a person and you killed that person in either case, but the interpretation is what puts you in prison for much longer or shorter time. So, interpretation really matters a lot, and the question is how we can be sure to get this interpretation right.

That, of course, depends on what we understand mental states to be, and here there is a long history of different proposals. Going back to the beginning of this lecture, it all starts with the common-sense or folk-psychological understanding of mental states. We all believe that we have mental states because we at least occasionally are conscious of our own mental states; we observe ourselves having feelings, beliefs, and desires, or lacking beliefs, desires or feelings. This is what we call *introspection*. It is a naive and uncontrolled introspection because we cannot really find a systematic way in which we can introspect. However, this experience carries the intuition that we really have, and that there really are, mental states of some form.

The nineteenth century psychologists, first and foremost the German psychologist Wilhelm Wundt, tried to systematize this notion of introspection by arguing that one could train one's own conscious thoughts and feelings in such a way as to observe one's mental states under controlled conditions. Wundt sat up a number of routines on how to perform this systematic and controlled experimental introspective psychology. However, at the beginning of the twentieth century there was a backlash, in particular from American psychologists who argued that psychology is really only in the business of observing public events. That is, motor behaviors of an individual. Only those can be objectively observed while thoughts and feelings, which may or may not exist, are not part of the science of behavior, hence they should be put aside. So what now becomes important is to develop regularities according to these methodological behaviorists, in order to discover and develop regularities between stimuli and responses. Here, of course, many of you know the famous experiment of classical conditioning performed by Pavlov and his dog. The bell that is rung whenever the dog is being given food, and the dog salivating first because it gets food but later on when only the bell is rung and no food is given, the dog still salivates in response to the bell-ringing stimulus. Such an observation speaks in favor of the idea that we can generate regular associations between originally neutral stimuli. Therefore, this constitutes an example of when we have a regular relationship between such publically observable stimulus and a publically behavioral response without talking about any inside thinking or feeling mechanism in the dog.

Methodological behaviorism in the first half of the twentieth century was very influential in psychology and continues to be so in some ways. However, in the late nineteenth-fifties and the early sixties there was a

countermovement. People, first and foremost the American linguist Noam Chomsky, argued that methodological behaviorism could not explain how humans can know so much even though they have relatively little input. Chomsky in particular developed this argument based on the example of children learning languages. He argued that humans must have some kind of innate learning mechanism that processes inputs in a very effective way, and that these mechanisms must be domain specific and innate. Further, that science should in fact investigate cognitive internal mechanism instead of only trying to do a mapping between stimuli and responses.

From this, amongst others, developed what we today call cognitive science. Cognitive science, amongst others, holds that theories refer to mental states, but only if those mental states are properly operationalized with respect to observational variables. So it is not a return to the old introspective psychology. Instead, introspection is generally not admissible and we have to connect the interpretation of people's mental states and cognitive processes in the light of publically observable behavioral evidence.

This raises an interesting question. Here is a case that I want you to think of. Think of a computer versus a human in a chess tournament, for example the computer Deep Blue (by now a very old computer, which does not operate in the same way as modern computers). Consider the claim "Deep Blue wanted to lure Kasparov's (a Russian chess player) queen by offering a pawn". What makes this claim true or false? Again, I want you stop reading and try to answer the question before reading further.

Whether this claim is true or false depends in the first place on the behavior of Deep Blue and Kasparov. But secondly, it also depends on what kind of theory of mental states we have. The old introspectionistic psychology held that only conscious animals introspect and thus observe their mental states. Then, only those animals actually have mental states, which means that they are largely restricted to humans. So from an introspective perspective, machines and non-human animals cannot be attributed mental states. Methodological behaviorists, like Watson, denied that mental states have any scientific relevance whatsoever, so behaviorists would have probably constructed pretty complex stimulus maps for Deep Blue and denied that Deep Blue was in need of any desire at all. However, cognitive scientists, like Chomsky, might have argued that Deep Blue's behaviors is sufficiently complex that it indicates some innate reasoning process, thus licensing attributing of mental states.

Now, maybe cognitive scientists would be wrong about Deep Blue. After all, Deep Blue was programmed with a complex set of behavioral strategies that perhaps could be mapped in way as the methodological behaviorists suggested. For example, "if your opponent moves in a certain way A then react with the response B" is perhaps something that would be of the form of such a stimulus-response schedule. However, when we look at the current chess computers, like AlphaZero, they do not

consist of such a long list of conditional response strategies. Instead, they have been programmed with a general learning algorithm and they do the learning themselves, which is really amazing. In this example of AlphaZero, the computer has become so good at its task that it has developed its own cognitive mechanisms, and it is indeed the question of what these cognitive mechanisms really are. For that, we would need to study the mental states of such a computer.

Incidentally, I should say that the cognitive revolution that I have mentioned before when I mentioned Chomsky was also promoted through an analogy of human thought and the computational functions of computers. It was computers who gave many cognitive scientists the relevant analogy and they conceptualized mental states as the counter parts of computational states or functions in computer, like memory storage and retrieval and so on. So for today's cognitive scientists it is not a problem at all to attribute a belief or desire to a machine in this way, if the behavior is sufficiently complex that it cannot be otherwise explained.

That leads us to what the philosopher Daniel Dennett has called the **intentional stance**. We assume an intentional stance towards particular agents. In the first place, those are humans, but they can also be animals. For example like your pet dog or maybe some animal that you are afraid of which you start talking to in a way where you attribute mental states to it, but you also do this to machines. So what you do when you assume the intentional stance is that you treat the object whose behavior is to be predicted as a rational agent. You figure out what beliefs ought to have given its space in the world and its purpose. As well, you figure out what desires it ought to have in a similar way. Then you decide based on these beliefs and assigned desires what the agent ought to do in this case. Once you concluded that – it is basically an inference of practical rationality – you infer from that what the agent actually will do. In other words, you generate a prediction. This is how interpretation works.

The intentional stance: Attributing mental states like beliefs and desires to an agent A (a human, an animal or even a machine), assuming A to be rational, and using this to predict the actions of A.

9.3 Quality criteria for interpretative methods

Now that we have clarified what an interpretation is, we can discuss the quality criteria – the methodological standards for such interpretative methods. Interpretation faces at least one main problem, which is quite obviously seen from our example of the usability experiment. The researcher observes one input (one instruction) and one behavioral response, but makes a number of (actually quite many) mental-state assignments in order to make this single instance of behavior understandable. We have understanding, beliefs, desires, and presumably multiple of them, assigned in order to understand this particular behavioral relationship. How does that work? There is obviously a multiplicity of possible, plausible assignments. We can call them alternative mental-state models. There does not seem to be any criteria for choosing between these. They all seem to be feasible. You might just pick out one, and if that one does not suit you anymore, you might choose another one. This obviously invites the criticism of an *ad-hoc* modification, where an explanation just can be suited to one's non-

epistemic needs of just have some ‘just so story’ without actually satisfying any methodological standards for scientific theory. So far the criticism.

Of course, the first step is to note that interpretation never happens just of a single instance of behavior. That is why we often describe not just individual beliefs, but rather personality traits of various sorts. *He is someone who likes this. She is someone who has the following type of beliefs.* This is supposed to be across time and across different contexts, so that individuals become these collections of stable mental states that we can identify time and again. Of course, from a scientific perspective that does not help us that much, because the social scientist who interprets a subject might not be particularly familiar with that subject. Hence, we need to find other criteria or ways around this problem.

The first strategy might be to *avoid interpretation altogether*. After all, why not be a methodological behaviorist where you can! Remember that taking the intentional stance is a choice. It is a particular view of an object, subject or system that we are taking, and there are reasons for this – typically complexity of behavior sufficiently high that we cannot achieve an explanation without attributing mental states or mental processes. However, if—inversely—the complexity is sufficiently low, then we might well avoid mental-state assignments altogether. Instead, we might investigate, like the behaviorist suggested, stimulus-response relations in individuals. Behaviorism, after all, is not entirely rejected; it is still a viable scientific strategy in many domains, and where we *can* apply it, there is no reason not to do that. We might also try to avoid the interpretation or systematization of individual behavior altogether. Sometimes we can move on to a social level, investigating relations between social rather than individual properties. This will be discussed later, at least for the social-science students, in the section on holism. Lastly, we might investigate relations between physiological properties, for example brain states or hormone levels, and behavior, instead of interpreting behavior. This would be another way to get around mental states. However, it should be said that much hyped neuroscientific research on behavior currently faces so many methodological problems of its own that it might not offer itself as an obvious replacement here.

Anyway, this is one set of strategies how to get around some of the methodological problems of interpretation. However, instead of avoiding an interpretation, we can try to improve it. Improving, in the first place, involves being very clear about the scope of the interpretation. We can try to sample individuals in a way that minimizes the interpretation needed. For example, in a usability experiment, we only select experimental subjects who are competent users, that is, they have certain beliefs about how things work, and those beliefs are largely correct. We might also select them so that they have, for example, positive attitudes towards the use of this kind of device, thus avoiding any desire to sabotage or not really participate in an experiment.

A second strategy for improving interpretation is to look for more evidence, beyond the instructional input and the behavioral output, for uncovering mental states and processes. For example, when dealing with someone working on a computer screen, it is very helpful to use eye tracking. Eye tracking gives us evidence. Not evidence *independently* to explain behavior, but evidence *for* attribution of certain mental states. Someone whose eyes flit about the screen, repeatedly searching for something, expresses bewilderment or confusion, while someone who is very focused, looking at one point, typically shows that there is a clear intention and a basis understanding.

Another alternative, as already mentioned, is a thinking-aloud protocol. Here, again, we are generating evidence that helps us to narrow down the possible mental states that we can attribute. Now, note that this is different from avoiding interpretation. Now, we are using evidence in order to select the right mental states, but we are not avoiding this altogether. Hence, it is a different strategy than, for example, neuroscientific or holistic approaches.

We also improve interpretation by making use of researcher – a competent observer. 'Competent' is here meant in at least two senses. First, in the sense that she is competent with and knows the device or the task that the subject is using or performing. Secondly, that she knows how to observe, and is aware of the additional evidence for, mental processes, which might express themselves in the subject, for example through fidgeting, sweating, being sleepy, expressing anger, and so on. Those are very important features that a competent researcher will note, and will therefore record in the research documents, in order to make interpretation more precise.

We should also include checks for coherence in an interpretation. In the first place, this involves *attention checks*. In a survey, it is a good idea to at some point just give a task that tells the subject to, say, move the cursor to some exact position. Someone who is just breezing through a survey, just clicking randomly at things can be caught by such attention checks. (Because, after all, most survey participants are paid, and they are not paid for the quality of their responses, rather just for participating. Hence, there is a real danger that someone wants to speed up, not thinking about the answers, but rather just 'go through them'.)

We can also *repeat questions*. In slight variation, a question that we asked early on in a survey or an interview is repeated at some later point. If the subject answers in a fundamentally different way, then we can conclude that the subject either did not pay attention or did not understand the question, and therefore we might have reasons to exclude this particular data point from our research.

We can also engage in *respondent validation*. In some cases, it is useful to afterwards—say, after a task has been satisfied or an interview has been performed—allow the subject to go through this again and check whether these were indeed the answers or choices that they wanted to make. That might correct inattention and it might help. However, it might also introduce further observer effects, so one needs to be careful here.

Finally, one might use *intersubjective checks*. Until now, I have assumed implicitly that the researcher who interprets the data is just one. So you are the subject, I am the interpreter, and I am just acting on my own competences here. However, there is no reason why I should not aim to, for example, systematically perform interpretation in collaboration with a colleague. Both she and I look at your responses, and we see whether we agree and interpret your responses in the same way. This leads over to the court case discussed earlier. In the court, interpretations are very important. That is why in many court systems, we have a jury system, so it is a number of people, all of whom need to be convinced that a particular interpretation is the right one. Each jury member observes the available evidence independently; so that we have multiple interpretations that need to converge to the same result. Something like this is very useful to be applied in the sciences as well.

To conclude, clearly interpretation is a fallible method, but so are all other inductive methods in the sciences, and not just in the social sciences. Interpretation is also widely practiced outside of the sciences, and often leading to very weighty decisions, as the example of the criminal court shows. Interpretation studies can be more or less reliable. We can make them

more reliable by taking into account certain quality criteria. Therefore, the more reliable ones produce more valid data, and the more systematically we employ the strategies mentioned here, the more confident we can be in the validity of these interpretations.

9.4 Qualitative Methods

The social sciences often use qualitative methods. I already mentioned that this has sometimes lead to the attempt to fundamentally separate the social from the natural sciences, saying that the social sciences are fundamentally qualitative in some not very precisely defined way. However, even if one does not want to characterize them generally as qualitative, surely many of the methods, or many of the data that they use, are qualitative.

Take for example the recent SINUS youth study, published this year in Germany, that uses narrative interviews, open questionnaires and photographic evidence in order to explore adolescents' attitudes about growing up: about their perspective in life, about their assessment of the current situation and so on. The methods they use and the data they generate are not quantitative, in the sense that they do not much involve numerical representation. Instead, it is typically in words and also in recordings and photographs. The way the study is analyzed is also qualitative, in the sense that there is no regression analysis; there is no attempt to model in a quantitative way. Instead, they propose seven broad categories, into which they sort their different individual subjects, ranging from conservative bourgeois, to the precarious, to the social-ecological, to the experimental-hedonist and so on. They have some criteria for why they are sorting individual people into the respective categories, but these are deliberately left somewhat vague and overlapping. Furthermore, the conclusions that this study draws are also qualitative, as you can see in this quotation.

> "In most mileus, secure living conditions are more important than status, success and advancement. Many young people wish to arrive 'at the center of society', material desires and goals are relativised…. The hedonistic mentality that was once so typical of young people continues to decline: partying, fun and action are becoming less important." SINUS Youth Study 2020

This is certainly interesting for someone being in their teens or early twenties today. It is also interesting for those who are trying to sell all kinds of services or design all kinds of devices for this cohort. But it is in a way quite 'fluffy'. It does not give us precise numbers. It talks about comparisons – something is 'more than', 'they are more important', something is 'in decline' – but we are not told to what extent: it is more testing and reporting on an atmosphere than giving a precise diagnosis.

9.4.1 Are qualitative and quantitative data types fundamentally different?

So one question I want to discuss now: is there a fundamental difference between qualitative and quantitative data types and methods of analysis, or are these exchangeable approaches? For this, let me draw an analogy that has nothing to do with the social sciences, going to a very simple, perhaps pre-scientific way of dealing with objects. Imagine that I have three stones that I am passing around to you; so you have them in front of your screen and can touch them, lift them, throw them, roll them, or whatever you want. Then, I am asking you to describe what you are experiencing. Well, typically what you will generate is a text (you probably just give a verbal report). You might say, "out of these three stones, *this one* is the heaviest one, *this one* is the smoothest, *this one* has a slightly darker color" and so on. So you are comparing, but you

are not assigning numerical comparisons. Either because those are not immediately obvious how to perform if you only have your hands and your sense organs at the ready.

However, of course you could now go and say "I want to know more precisely how much heavier, say, stone number one is than stone number two, in comparison to stone number three. Then, you need to develop an operationalization. You might need to construct some form of scale in order to make this comparison in a way that allows an interval, or even a ratio comparison. The important point here is that really the move from the qualitative to the quantitative data is just a refinement of information that is, at least partly, already contained in the qualitative data. The qualitative data allows just for ordinal rankings of weight, smoothness, shade and many other properties, and for each of these we can now develop quantification, if we so desire

In the social sciences, we often *can* similarly quantify qualitative data. Sometimes, data already comes quantified after all: when we record prices, demographics or votes, then we are already making use of quantitative representations. Then, there are ways how to generate quantitative data from objects or properties that we may intuitively think are qualitative in nature – for example desires or beliefs. There is a formal theory that allows us to generate an interval-scale comparison between preferences, which requires that we go through a particular procedure of offering to our experimental subjects – particular choices between lotteries. This allows us to make comparisons beyond the ordinal – to actually assign the comparative differences between the intensity of desires. Similarly, we can do that for beliefs.

Are qualitative and quantitative data types fundamentally different?

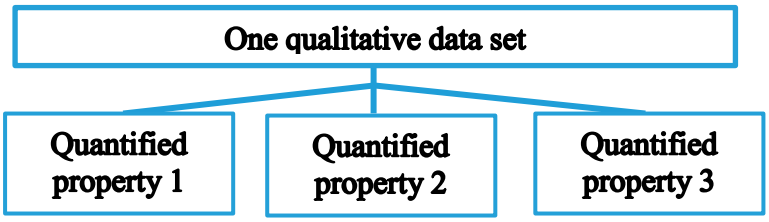
- Data types share at least some informational content
- Qualitative data often can be quantified

So there are standardized procedures how to quantify certain properties in social sciences. In other cases, we might not have a *standardized* procedure, but we nevertheless have strategies how to do so. We can take a large set of verbal recordings from an interview and code for particular keywords. Then we can, for example, compare how often the subject used positively evaluative versus negatively evaluative adjectives in the interviews. That might help us to systematize the analysis of such interviews – particularly if the interview volume is large.

We can now see that the distinction between qualitative and quantitative is much more permeable than people would have it. Qualitative data is the material that can be collectively analyzed in its own right, and sometimes that is exactly what we want. The SINUS youth study, for example, makes very clear that they want to provide readers with a lot of the unedited interview material that they gathered from these adolescents, and it is important to them that they don't only give them summarized, quantitative amounts. So qualitative data is important in its own right, because it helps us to grasp something in more depth and get a better 'feel' for certain cases.

Qualitative data can also be developed further into quantitative data. Then it becomes the raw material: an intermediate on the way to the final inferences. Often, of course, we are making use of both. We are using the qualitative data in its own right and are also quantifying some of it, and that is what gets us this **mixed-methods** toolbox. A word of warning here: it might sound as if quantitative is always better than qualitative, but I am far from saying that. In fact, the qualitative data has its own very

Mixed method (of data types): Using qualitative data both in its own right and as raw material for generating quantitative data of one or several properties.



Mixed method

important purposes. It just does not contrast in the way that some people want. Furthermore, sometimes qualitative data is the best that can be collected in a given environment. For example, it is questionable whether it is always worthwhile quantifying (or asking for quantified) data from our subjects, in particular if those subjects do not really understand what we are asking them.

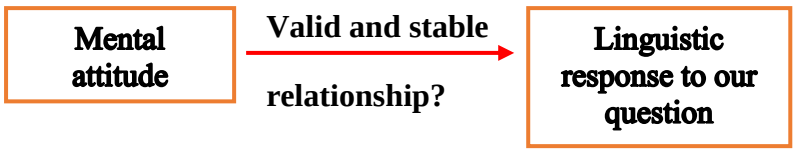
Think of the common Likert scale. It can be presented in different ways, and some people have argued – I think successfully so – that it is not the best way, but sometimes you still find it. You simply have questions, and the people are supposed to answer with answer options 1 to 5. That might be *less intense* to *more intense* or *disagree* to *agree*, or something of that sort. Now, given that we already number them 1 to 5, we might now assume that the subjects understand the distance between those five answer options to be the same. That is, the distance between 1 and 2 is supposed to be the same as between 2 and 3 in intensity, and the intensity of the minimum and the maximum is supposed to be equal across all the different questions. If that is true, then we can, first of all, calculate the mean of different people answering the same question, and we might even sometimes aggregate the answers to different questions along some broader construct. But what if the subject did not understand or interpret the answers according to these assumptions? What if they interpreted 1 to 5 in a different way? Simply in some qualitative sense, maybe just as an ordering of some sort? If the subject answers in that interpretation, it would be wrong to make use of the Likert scale as if it were measuring attitudes on an interval scale, so here is a word of warning against unnecessarily or mindlessly quantifying.

Quantification also often narrows the focus. Go back to our three stones. If I were to ask you (which I have done in the classroom) what sort of properties you are observing, someone usually comes up with an altogether new property that they are observing in these stones. Something interesting, exotic, an unusual perspective on them. If I start out by asking you to quantify certain properties, then I am focusing you on these properties, and that narrows what we can extract from your responses. So what we often want, let us say in a focus group, is that

people can freely express their evaluative or attitudinal perspective. We again can gather that into a qualitative data set and then we might identify a number of separable properties, each of which we might then proceed to quantify. So here, you see another rationale for a mixed method that first focuses on the qualitative and then proceeds to the quantitative.

9.4.2 Construct validity

When we are asking interview questions, like "are you nervous when doing this", or "how much did you enjoy doing that", we are assuming that subjects understand what we mean with the terms we use in the question. We also assume that we observe the relevant properties through the lexical interpretation of the subjects' linguistic responses. So, the subject reports that they have been *very nervous*, *quite nervous* or *not so very much nervous*, and we are supposed to interpret that in a particular way. Now, if you think about examples like a subject being embarrassed, shy, annoyed or secretive – or subjects having different linguistic



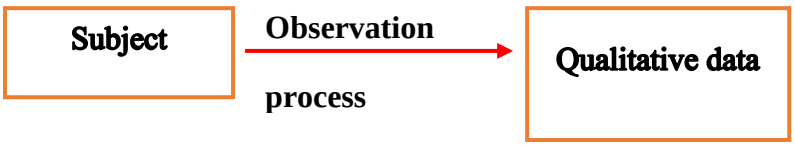
understanding or competences, perhaps due to differences in culture and education – those assumptions are often not legitimate and justified. Thus, when we are observing responses to our questions, we are assuming that there is the appropriate relationship between the property that we are interested in finding out – some attitude that is not directly observable – and the linguistic response that we are recording. However, sometimes, this appropriate relationship is not given. This brings us back to the issue of operationalization. We operationalize our measurement of attitudes, even when we measure only qualitatively, by assuming that there is such an appropriate relationship. But that often is not given, and if not, then of course our measurement does not have construct validity.

I can point you back to the earlier chapter on measurement, where I discuss conditions of good operationalization. I will not repeat them here, but what I really want to stress is that when we discussed quantitative data, it was obvious that we needed to have some operationalization, because we could not do without. It was clear that we had to have some unit of comparison, and therefore we needed to have the operationalization on the table. When we do qualitative research, we often do not make that so explicit, but that is just negligence. It does not mean that there is not a question of operationalization. To the contrary, there is the issue of operationalization in qualitative research just as much, and we need to address that. These are important quality criteria. Can we give reasons for why we believe that, say, a particular linguistic response is indeed connected to a particular mental attitude or mental state of the individual? If we cannot, then we have to look further into the operationalization.

9.4.3 Can the validity of qualitative data be evaluated?

Some people fundamentally question whether qualitative data can be evaluated as valid. I do not really see the problem. We have a subject, the subject will be in some state or other, and we are performing an observation that generates some data from which we infer the state that the subject is in. So an evaluative question about the validity is: is the information contained in the qualitative data accurate, precise and stable about this subject?

Here are a number of quality criteria for qualitative methods. First, we want to ensure that someone who wants to check on how we perform this study, and might want to repeat it to see whether it is reproducible, can **repeat** the study. So, document the techniques of data collections that you performed, and make your modes of data analysis transparent



(whatever those might be). Then, be mindful that you as the interviewer, or even as the designer of the survey, will likely have some **observer effect**. We cannot control for this, because in surveys and interviews the only intervention is the question. Thus, we cannot really generate a control group where we could switch that off. Instead, we might aim to minimize (ideally eliminate) the observer effect by remaining as neutral as possible (whatever that might mean). However, what we can do is to reflect on our **observer biases**: we describe the way in which we think that we have affected the subjects – what the relation between the researcher and the subjects has been. This might include describing the researchers' sociocultural positions and properties. Last, in order to improve the **robustness** of a study, it is worthwhile, just as for quantitative data, to aim at **triangulation**. So, we are trying to apply multiple methods of observation about the same properties, and we are interested in individuals' preferences over a particular domain of objects. Then, we might want to both ask them for their subjective evaluations, so that they state their preferences, and also put them in a situation where they are given a budget and they are choosing from that constrained budget which objects to buy, so that we can actually get behavioral evidence: a choice based on such constraints. These are surely not all, but at least some, quality criteria, by which we can assess the goodness of qualitative methods.

Ensure repeatability – Procedural rigor: Document techniques of data collection, make forms of data analysis transparent.

Control observer effects: Since questioning in interview/survey is the only intervention, control group and blinding cannot be used. Instead, minimize/eliminate effect of observer on subject.

Control observer bias – Reflexivity: Describe the influence of relation between researcher, topic & subjects. Describe researchers' sociocultural position.

Robustness – Triangulation: Use multiple methods of observation, e.g. combine revealed and stated preference analysis.

9.5 Case studies

9.5.1 What are case studies?

Social scientists often use **case studies**. Case studies are not the same category as qualitative methods or interpretations, and therefore need to be discussed separately. We characterize them by their small number of cases, often just one, that are studied; by the individual cases being chosen for some specific reason, so that each case has its own identity in some important sense; and by all cases typically displaying the same outcome, so there is no variation in the outcomes. Case studies typically do not involve experimental manipulation, and we can say that the goal of case studies is to understand a specific case.

Case study: A study characterized by using a small number of cases that are chosen for some specific reason and display the same outcome. The study typically does not involve experimental manipulations. The goal of the study is to understand a specific case.

This contrasts with what we might call **multivariate studies**. Here, we are looking at a large sample from a defined population, randomly or in some other representative manner, so that the sample is representative of the population. We also want the values of the so-called *dependent variable* to vary across different observational points. Contrast this to the cases, which all display the same outcome. Multivariate studies either involve, or often try to mimic experimental manipulation – think of natural experiments – and the goal is to identify and understand general patterns that we find in a population, rather than just looking at specific cases.

Multivariate study: A study using a large number of cases, chosen from population randomly or in other ways representatively. The Values of the dependent variable vary across observation points. The study involves or tries to mimic experimental manipulation. The goal of the study is to identify and understand general patterns characterizing a population.

It would be a misconception to think of case studies as only being qualitative. We often measure with numbers in case studies, so that is not part of the distinction, and case studies don't need to be of small units; we might consider the case of an individual, but we also might consider cases of organizations, large-scale events (e.g. the dot-com crash in the international stock exchange around year 2000) and so on. So we have any unit of analysis that might serve as a case.

9.5.2 Why perform case studies?

There is a number of important motivations for performing case studies. First, maybe simplest, if we are operating from a *falsificationist* perspective, a single case is enough. We identify a falsifying case that rejects our theory. Now, we develop a new theory until that hits another, again, potentially single case. So, here the case actually has quite a lot of effect. In other situations, we are interested in a phenomenon that is sufficiently complex for us not to know, at least at this point and with our background knowledge, how to approach it, otherwise than by doing an in-depth study. Often this involves collecting evidence for feelings and motives of actors, and in those cases, we might end up with a so-called *analytical narrative* of this case (this and the following concepts will be further explained in the subsections below). In other situations, we might want to compare a number of cases without knowing exactly what the underlying population is. We might be interested in learning about necessary causes (note: not sufficient, but necessary), and those are then helpful to get a better understanding of the population boundary, that can later on be applied in multivariate analysis. Such case approaches are often called *qualitative comparative analyses*. Finally, again thinking

about causes, we might be interested in the cause-effect relationship in an individual case; or we might be interested in how an intervention, a cause, actually operates on an effect: what sort of mechanism is at work in the background. Those questions cannot be answered by population studies or multivariate studies. Instead, we need to do an in-depth case study, in this case called *process tracing*, in order to learn more about these mechanisms.

9.5.2.1 Analytical narratives

Let us start with the **analytical narrative**. In this case, we build a model to express the logic of a possible explanation for a given phenomenon, typically in the context of a unique case. For example, '*why did Chinese early adopters in 2015 hesitate purchasing the Apple watch?'* What happened there? We might give a narrative that, in the first place, makes explicit our research question. Then, we select a number of cases, in this case a number of early adopters who have been identified, maybe in some other study, as being early adopters, and use them for the study. We then offer a narrative that elucidates these people's preferences, their key decision points and the various possibilities they had when making a decision. Then, at the end, we try to test our narrative by drawing out further implications and see whether they accurately predict features of this case. So, we might give a narrative of the sort "Well, the Chinese haven't taken to personal-health related consumer technology in the same way as Americans have, and that has made them more hesitant with regard to this particular device". One way in which we can improve such analytic narrative is by considering alternatives: alternative narratives, alternative models for the same case. What are the alternative explanations? Was it maybe instead the austerity policy introduced by the Chinese government around the same time, that made people think that it was wise to not show something that looked very 'flashy' and was visible to everybody? Here, you can then draw out a number of implications that each of these models would have, and test them by seeing whether the case studies you performed—that is, the individual early adopters who you investigated—actually show one or the other these implications.

Analytical narrative: The method involves (1) making explicit the research question; (2) selecting cases; (3) using narrative to elucidate the principal players, the inferences, the key decision points and possibilities; (4) evaluating the model through the testable implications it generates; and (5) considering alternative narratives.

9.5.2.2 Qualitative comparative analysis

Analytic narratives focus on one or a small number of cases, but sometimes we want to expand into comparing systematically a slightly larger of cases, in particular if we are interested in potential necessary causes of phenomena. Here is an example. Think of the question *'what are the causes for successful labor strikes?'* Here is a table that for each role gives us a number of cases in the right column, where certain causal factors were either present or absent, and where the successful strike (the outcome) was achieved or not. We have three potential causes: (A) booming property market, (B) threat of sympathy strikes, and (C) large strike funds. Those are now three potential causes, but the question is, "Which of these are necessary for a successful strike?" Well, we can find out by comparing all those cases where the outcome is a successful strike. We do not need to actually compare the other cases at all; they are only presented for completeness sake and will not be made use of.

In the first step, we establish a link between these causes (or combination of causes) and the successful outcome. What we do is to look at cases where both the outcome and the cause is present, and then we get the four colored rows in the table. From those, we can get four combinations: **A, B and C** in the first row; **A, B and not C** in the second; **A, not B and C** in the third; and **not A, B and not C** in the sixth row. All of these cause a successful strike.

A	**B**	**C**	**S**	**Nr of cases**
1	1	1	1	3
1	1	0	1	2
1	0	1	1	6
1	0	0	0	9
0	1	1	0	3
0	1	0	1	5
0	0	1	0	6
0	0	0	0	0

In the second step, we can remove some redundancies. We do that by comparing pairs that differ in only one cause. We can compare the first and the second row and see that they only vary in the cause C, so really, it is A and B that cause the strike, irrespective of whether C is switched on or off. Thus, we can already say that **A and B** is one of the causes. A second comparison is between the first and the third line, where the only difference is B, so here it does not matter whether B is switched on or off, and so **A and C** is another combination that causes S. Finally, we can compare the second and the sixth row, where the only difference is A; here, C is switched off, B is on, and it doesn't matter whether A is switched on or off, so it is **B and not C** that is another combination that causes S.

Now, in a third step, we are comparing whether there are cases that are successful without some of these factors present. What you see is that there are some doublings, so in some cases, in particular in the third and sixth row, there are two factors absent. However, it turns out in this comparison that the combination **A and B** is nowhere needed. We can always either have the factor **A and C** present, or the factor **B and not C** present. From this, we can conclude two things. First, there is not a single factor combination that is necessary for all cases in S; we need some disjunction. However, we do not need the disjunction of all the three factor combinations, but instead, it is enough—all cases are covered—if either **A and C** is present, or **B and not C** is present. **A and B** is a redundant factor combination. We can conclude that the potential necessary causes of successful strikes in this case (which, admittedly, is a bit of a toy example that allows to very clearly bring this out) are necessary for the success of a strike.

Note that, in the previous section, we have only compared cases where the strike is successful. We did not even have to compare or investigate the cases where the strike was not successful,

so the method of case study only concerns cases of the same outcome. In multivariate analysis, we cannot make this distinction between necessary and sufficient causes. Rather, multivariate analysis presupposes that we have already identified the necessary causes of the phenomenon. Therefore, case studies in the form of **qualitative comparative analysis** can be seen as an important step towards developing multivariate studies, so here we have another example of where case studies play an independent and very important role.

Qualitative comparative analysis: Analysis of cases performed in order to learn about necessary causes and thus better understand the boundaries of the underlying population. The results can be applied in a multivariate analysis.

9.5.2.3 Process tracing

Finally, **process tracing** is another type of important case studies. In process tracing, we are determining how cause and effect relate in a single case. This is something that cannot be done in a population study. Population studies, like multivariate analyses, discover the mean influence of a cause on an effect across all the cases in the population. Thus, in the first place, such a population study does not reveal *how* the cause influences the effect; it just says that the cause influences it to a certain magnitude. Secondly, if there are many different mechanisms in the population, then we cannot draw a conclusion from this population-effect-size relationship to any other individual case. Simplest is to think of a case of two patients that both take a pill. The one has no reaction to the drug, while the other patient improves 100% (whatever that might mean). Now, the mean effect of this drug, given that we only have these two people, would be 50%, but of course, that is entirely irrelevant for the study of individual cases. In the individual cases, it was either 0% or 100%, not 50%, and once we have a large population that we study in this way, and we have more heterogeneity of mechanisms, then we have quite a problem. We cannot infer from the population study to the individual cases. Instead, what we need is to do a case study of the individual cases and trace the mechanisms in those cases.

Process tracing: Cause-effect relations in a single case **vs.** Cause-effect relations as mean effect size in the population

Four steps

(1) Formulate hypotheses about alternative mechanisms

(2) Work out differences in empirical implications

(3) Check case data for these mechanistic fingerprints

(4) Conclude which mechanism it is

Here is a very simple example of two clocks that operate quite similarly, but there is an important distinction. They are driven by mechanisms, which differ in one aspect: one clock is driven by a mainspring, and the other clock is driven by a pulley. They both show similar cause-effect relations, so you cannot differentiate them on that, but you can now start investigating the individual case in the following four steps. First, you *formulate a mechanistic hypothesis*: what are the mechanisms that could drive this? Then, you work out what sort of *empirical implications the differences* between these two mechanisms would have; one simple example is that in the mainspring clock the spring force (torque) continuously lessens as it unwinds, while in the pulley it does not. Now, we can *look in the case for such mechanistic fingerprints*: is there such empirical data, in the case that we are studying, that shows this continuous lessening of torque, or does it not? Based on that, we might then *conclude which mechanism it is*. In this case, we did not perform any population study. We only looked at one case. We did not intervene in the case at all. Instead, we are just inferring from these mechanistic fingerprints to the operation of a particular mechanism.

9.5.3 How to select cases?

In the previous section, we saw a number of examples of why we might choose case studies over multivariate analysis and how they can be important; but of course, as I said, a case needs to be chosen for the purposes at hand. Thus, we have a selection problem. We need to give reasons for why we are choosing particular cases and not others. Very broadly, we can have either representative sampling, which is more closely associated with multivariate studies, or we can have more information-oriented sampling, so we are picking out individual cases because they respond to our research questions in a particular way: we have particular reasons for choosing them. Such reasons could be, for example, that a case is particularly extreme, and thus presents the phenomenon in relatively pure form. It could also be that the case challenges an important theory, is very typical, or is historically and culturally significant—for example, "this is an early adopter: she is a pioneer"—something of that sort. Therefore, we chose *this* person, or *this* particular event as our case study. The reason to why we are choosing this case should be connected to the background theory. That might sometimes be implicit, and it is important to make that explicit: to make clear *why* we are choosing this case.

9.5.4 How to generalize from case studies?

Another problem for case studies, of course, is that we study the case, we learn about the case, but we have little reason to transfer what we learn about the case to anything else. That is often seen as a substantial disadvantage by defenders of multivariate analysis, which of course it is. However, we should in the first place say that the case study aims to answer different questions than a multivariate study, so it is unfair to measure the importance of a case study by its inability to be transferable. A second answer to this criticism is that sometimes such cases can be improved upon to such an extent that they might become transferable. For example in an analytic narrative, we can try to investigate whether the narrative has further implications for this case, so that we might be able to predict things that we have not observed yet, and then go and make more observations about this particular case. If we can do that, it can strengthen our confidence that it is actually a good model of this case, and therefore might also strengthen our confidence that it is a good model for other cases. Furthermore, like with any model, we can of course test case studies by out-of-sample tests. We can perform another case study with a substantially different case, and then see whether the model that applied to the first case also applies to this case. If it does, then that gives us some confidence that in fact the case study can be transferred.

Ways to improve upon transferability

- Does the model yield other predictions that can be tested on this case?
- Out-of-sample tests: perform another case study and check whether model from first study applies

9.5.5 Summary: How to improve case studies

To conclude, I argued that case studies can be improved in a number of ways and that they are subject to a number of methodological considerations. Here is a summary of these criteria. The first one is to *specify the research question*: why would we want to perform a case study, rather than a multivariate study? That already shows, or should show, in the research question, as these studies tend to answer different questions. Then, *specify the theory background and the hypothesis* that you want to assess and, maybe not test, but evaluate in some form with the case study: what do you think are the driving forces of the phenomenon? Once we have that clearly specified, we can more clearly specify which cases we should actually select. We can also more specifically determine what we should observe about those cases. However, before we do that, we should also clarify *what constitutes a case unit*: is the case in some ways separate from context, does the case include some of the context? That is something that needs to be determined beforehand – before we actually perform the study. We then need *to make the case selection explicit*: what are the reasons for selecting this or these case(s)? Here, we should point back to both the research question and the theory background. In multiple cases, we should also be clear about *what makes these cases comparable*: do we select multiple cases from different contexts, or from the same context? That is, again, depending on the kind of questions we are aiming to answer. We next have to make sure that we *investigate all the implications of our theoretical hypothesis* for the case, and check whether the case satisfies them. That is crucially important in the process tracing, because that is what process tracing *consists in* – working out those mechanistic hypothesis implications, and then testing whether they apply in a particular case; but even in analytic narratives, we gain by showing that all implications of the theoretical hypothesis are actually borne out. Lastly, we should not forget to expose the case to rival theoretical hypotheses, and show that, in fact, our hypothesis (even though, of course, we cannot test it in a rigorous way) works better with this case than potential rivals do.

How to improve case studies

- Specify research question
- Specify theory background and hypothesis
- Clarify: what constitutes a case unit?
- Make case selection explicit
- If multiple cases: what makes cases comparable?
- Investigate all implications of your theoretical hypothesis
- Investigate rival theoretical hypotheses

9.6 Holism

So far, I have largely focused in social-science endeavors that focus on individual behavior, and in fact, a lot of social science pursues that path by endorsing the so-called **Methodological Individualism Principle**, which says that all good social theory is reducible to individualist accounts. Whatever social phenomena exists should always be reducible to the action and interaction of individuals.

The Methodological Individualism Principle: All good social theory is reducible to individualist accounts.

9.6.1 The ultimatum game

Here is a very simple example of such an interaction between individuals: the so-called ultimatum game. It consists of two players who have one choice each, and they choose one after another. Player 1 has $10 in her hands and can choose how much to share with Player 2. Once Player 1 has chosen and told Player 2 how much she is willing to share, then Player 2 also has a choice. Player 2 can either accept Player 1's offer, in which case the $10 are divided in the way as Player 1 proposed; or reject the offer, in which case neither player gets anything – the $10 are taken away again. Now, take a moment and think for yourself. What share would it be rational for Player 1 to offer, or if you prefer, more personally, if you were Player 1, how much would you offer in such a game?

It would be interesting to hear your answers, but unfortunately, that is not possible. I can tell you though, that according to the standard game theory, the only rational choice for Player 1 to make is to offer the smallest amount, that is, one cent, and not more. Why is that?[4] Well, there are two payoffs that Player 2 can receive, depending on his choice. He can either receive nothing (if he rejects Player 1's offer) or he can receive a positive amount of money: that what Player 1 has offered. Under the assumption that some money is always better than no money, it is rational for Player 2 to choose to accept the offer, irrespective if what the offer is, as long as it is positive; and, by definition, Player 1 needs to offer a positive sum, even though a very small one. Player 1 knows this, can think through Player 2's choices and can conclude that Player 2's rational choice is to accept, irrespective of how much Player 1 offers. Therefore, it is rational for Player 1 to offer the smallest amount.

Now, you might think "that is nonsense; that doesn't work", and most of you probably (at least when thinking, "What would *I* do if *I* were Player 1?" have chosen to share something much more. In fact, that is what we find when we investigate how people actually play this game. The ultimatum game is a wonderful model that has been used for laboratory behavioral experiments like no other model. All over the world, people have been offering these choices. Not just "what would you choose", but participants have actually been offered to participate in such experiments and go home with the payoffs that they win in this game. This has been played also in countries where wages are relatively small; so that one could make the sum much more significant than $10, say a week's salary. Nevertheless, people consistently show that, with some variation, they prefer a fifty-fifty distribution. People playing as Player 1 typically offer, with some variation, within the fifty-fifty region, and the acceptance rates for those who do not—who give some much more uneven offer—go drastically down.

Thus, we have an interesting contrast here between the prediction of a pure model and the assumptions of individual self-interest on one hand, and the actual behavior we observe when we let people play this game on the other. This raises the need for an explanation: why is there such a divergence? We have a number of suggestions that social science has provided. One is that there are cultural differences. Others have argued that it depends on whether Player 2 perceives Player 1 as having earned the $10, or whether she has just been given it, and also whether she was responsible for the offer, that is, whether she chose how to divide it, or whether it was some kind of random device. For example, if she chose to throw a die that would then determine the distribution, the unequal offers were more accepted. But perhaps the cultural—the vaguest of them all—is the most interesting explanation, because what we *do* find is that there are a couple of divergences. Even though by and large across the world, the fifty-fifty is

[4] See diagram in video (2:50).

dominant, there are some micro-societies where these divisions look quite different. In some such societies, sometimes Player 2 rejects offers that give them more than 50% of the share. In other societies, people systematically offer considerably less than 50%.

9.6.2 Social norms explanations

Overall, we can say that what matters are social norms, and we find that the fifty-fifty norm is a widespread, but not universal fairness norm that exists and expresses itself in most, but not everybody, playing the ultimatum game in this way. What I find particularly interesting here is that social norms are not individual properties. Of course, social norms are constituted by individual beliefs, but these beliefs are socially influenced and coordinated so that they have one direction. This directionality is not attributable to any individual – in a similar way as, say, language use cannot be attributed to the competence of an individual, but there is some grammar that develops, and no individual has developed that grammar.

Hence, we have, in social norms, a grammar of social behavior that does not seem to be reducible to, or at least not just constituted by, individual behavior. Therefore, explaining this divergence, or this behavioral phenomenon, of how people play the ultimatum game with reference to social norms *violates the Methodological Individualism Principle*. It is not a methodological-individualist explanation, because we are referring to a non-individual property. You might say, "in that case, why not reduce social norms to individual behavior then, in order to conform to methodological individualism again?" Here, there are two responses.

The first one is some version of what is called **ontological holism**. The argument is that social norms are things that really exist, but they do not exist on the individual level. Instead, they exist on some social level, and this social entity cannot be reduced to individual qualities. These ontological claims have run into various problems, so I do not want to discuss them any further. It certainly is a possibility, but the more interesting possibility is that we focus on **methodological holism**, which says that yes, social norms can be reduced to individual qualities; however, they can only be reduced on a case-by-case basis, that is, there is no general way how all social norms are generated, or constituted, by individual properties (be those beliefs or behaviors). Instead, we have a situation of *multiple realizability*, that is, in some societies the fairness norm might have come about through that society having undergone a long phase of merchant tradition, where trade, mutual trust and certain minimal fairness norms were important. Therefore, they evolved in the history of this merchant nation. In other cases, the fairness norm might have emerged as a byproduct of the religion that people practiced in that particular society. In yet other cases, perhaps in a third society, it has emerged as a consequence of particular political leaders who stressed and cultivated civic virtues including those of fairness.

Thus, we have different histories that might all have led to the existence of these norms – each of which, of course, consists of histories of individual interactions. However, these different histories do not map onto each other.

Ontological Holism: Social norms are things that exist, but they don't exist on the individual level. Instead, they are social entities, irreducible to individual qualities.

Methodological holism: Claims, in contrast to ontological holism, that social norms can be reduced to individual qualities, but only on a case-by-case basis. Social scientists aim at describing in general the relation between social norms and human behavior. A case-by-case relationship only gives us specific historical narratives, and we do not want to trade explanatory generality for methodological individualism.

Thus, the social norm that we used to explain the ultimatum-game behavior might be multiply realized through these different histories, and that is not very useful as the basis for a social-science explanation. Social scientists aim at providing a general explanation, so if there were a general relationship between social norm and individual properties, then we could maybe usefully reduce the existing social norms to these individual properties. We would then go back to a methodological individualism. However, the case-by-case relationship, with the possibility of multiple realizability, only gives us a number of alternative historical narratives, none of which can claim generality. That is not something that we want in our explanation. Therefore, it is preferable to give up methodological individualism, and instead stick with the simpler explanation that refers to a social entity based on such methodological considerations.

Part 10 – Economic methodology

To be added

Part 11 - Ethics

11.1 Morality and Ethics in Science

This chapter is about justifying moral questions and claims that scientists are required to take a stance on. For example: "May I perform experiments that might harm human subjects?", "Must I reject data if I suspect it to be falsified?", "What do I do if I suspect that my colleague is either falsifying data himself or is using such falsified data?", "Who should be counted as an author of a scientific manuscript?" or "What kind of research topics should I choose?" Answers to these questions can be of the following forms. You *may* do something, which means that you are morally **permitted** to do something. You *should* do something, meaning that you must do, you are **obliged** to do, something. Or, you *must not* do something, for example, you are **prohibited** from falsifying data. These forms alone do not characterize moral judgments, however. There are prescriptions that are not moral - for example, you *should* control background conditions when designing and performing an experiment, so as to increase the internal validity of your inferences. Such a prescription is methodological, not moral. So it seems that not all such forms, such normative statements, are moral.

Permission: Being morally allowed to do something.

Obligation: Being morally required to do something.

Prohibition: Being morally forbidden to do something.

What is morality then? We can distinguish two broad views. The **descriptive view** tells that morality consists in the code of conduct that are established in a society, a group, or among individuals. So, for example, the *Nuremberg Code* of legitimate medical experiments is such a code of conduct, which was written in response to the horrific human experiments performed in Nazi Germany before 1945. It sets limits of what is permissible and what is legitimate. That has been expanded and further developed in the *Declaration of Helsinki* in 1975. Today, we find many more such codes restricting experimenting with humans and with animals, but also those that address non-experimental questions of bioethics, for example, or questions of moral conduct for engineers. The descriptive view takes any such set of codes, any long lists that says what you are allowed to do, what you are disallowed to do, as an expression of morality, as long as it is accepted by *some* current people or groups.

Descriptive ethics: Descriptive ethics studies what people consider to be right and wrong.

However, there are two problems with this view. The first problem of the descriptive view is the existence of non-moral codes of conduct - there are codes of conduct that many people might agree with, but which intuitively are not moral codes. For example, *convention* is a code of conduct. In Sweden, we all agree that one should drive on the right side of the road. Yet, we would typically not think of that as a question of morality. You have misunderstood something about traffic if you violate that convention - but you have not acted immorally merely from driving on the left side, even if this results in producing a crash. *Etiquette* – for example, not to speak with your mouth full – is a prescription and violating it might be not pretty, but most of us would not think that that is a question of morality. The *law* requires a more interesting and more difficult distinction. But consider the many prescriptions, legal prescriptions relating to traffic behavior for example, that are enshrined in law. Those

prescriptions are typically not seen as moral prescriptions. On the other hand, there are many prescriptions that we consider moral prescriptions. For example, our particular obligations to our friends and how we can, but should not, violate them. That obligation has nothing to do with the law. At best, we have an overlap of moral prescriptions and the law, but we cannot say that the law is part of morality or morality is part of the law. Last, *methodology* belongs to this kind. As a scientist or engineer, you should use a particular method if your aim is to achieve a certain epistemic goal. That is a prescription, and such prescriptions are typically shared by a wider group of scientists, but they are intuitively not moral.

The first problem of the descriptive view is its inherent relativism. If morality is what some actual people accept as a code of conduct, there will be many conflicting moralities. One then cannot say that some behavior is morally wrong – one can only say that it is wrong with respect to code 1 (but perhaps not with respect to code 2, etc.). In the worst case, we would have to admit that even the Nazi scientists in Germany before 1945 performed their experiments according to *some* code of conduct that was realizing whatever ideals, however perverted they were, of the Nazi party. That seems entirely wrongheaded.

Therefore, not every established code of conduct is a code of morality. Some established codes of conduct might be immoral, and some moral norms might not be realized in any established code of conduct. In order to consider these possibilities, one needs a **normative**, not a descriptive account of morality. But how can one characterize morality normatively, without referring to established codes of practice?

Normative ethics: Normative ethics studies what people should do (irrespective of what people actually think).

My starting point is to define morality as that code of conduct which all rational persons, under certain specified conditions, would endorse (**universal rational endorsement condition**). The details of these conditions are controversial, but it is enough here to give a sketch of what is meant by that. In the first place, these are counterfactual conditions. If morality were defined as what actual people accepted, then we would be back at the descriptive account. Instead, the normative account imagines how an individual would decide, if his or her situation were substantially different from what it actually is. In particular, for any event E (let this be the action of a single person, or a government, or some other institution), the account assumes that (i) the individual has perfect information about how E affects any conceivable position a person might occupy in society or in any other group, and (ii) the individual does not know in which position she or he is personally in. This way, the individual will judge E not based on personal projects, or relations to some particular persons. Any instrumental reason for or against E based on one's self-interest, one's personal relationships, one's particular plans how to lead a life, all of these will not feature in the moral judgement of E, because by (ii) the individual is assumed not to know about them. Yet per (i) the individual will know about how E affects any possible position in society, and furthermore, she will consider the possibility that she might occupy that position.

Universal rational endorsement condition: The moral code that all rational persons, under certain specified conditions, would endorse.

Such a normative approach distinguishes morality from established codes of conduct. Established codes of conduct – conventions, the laws or methodology – typically are based on instrumental reasons, whose ultimate goals are not universally shared. They would not stand up to the universal rational endorsement condition, because people endorse them only due to their specific projects, specific positions, and specific conditions. Furthermore, bad codes of conduct that disadvantage some people can only do so because the position of these people in society is known. An individual would hardly rationally endorse such discriminatory practices is he could be in that position himself. Thus, the normative account, cashed out by the universal rational endorsement condition, can overcome the problems of the descriptive account of morality. The question then is, "What are the reasons that would make a rational person, or that would make you, endorse a particular moral code under these counterfactual conditions?" That leads us to the consideration of **ethics**. Ethics considers exactly these reasons. It is not just a selection or presentation of how one should behave. But rather, it investigates what reasons we have for endorsing certain claims that say, you must do this, you may do this, you should not do that. And in particular, in research ethics, then we ask, why should I, as a student, as a scientist, as a teacher, endorse certain codes of behavior? Or more generally, why is it rational for all scientists to endorse certain codes of behavior?

Moral vs. ethical / morals vs. ethics: One distinction between morals and ethics is that morals are properties of actions, intentions or decisions (being good or bad), where ethics is providing justification for *why* actions, intentions or decisions have this moral property. However, in many usages this distinction is not maintained, for instance the same behaviour might be described as immoral or as unethical, without any clear difference being intended.

The answer, these reasons, do not need to refer only to codes of behavior. There might be reasons for particular prescriptions, which later become part in the framework for a methodology, framework for reasons why to choose one method over another. So far in this course, I have only discussed *epistemic* reasons for why we choose one method over another – reasons based on considerations of scientific goals related to knowledge ("Episteme" in ancient Greek refers to knowledge, science, or understanding). But now, we might also think about the reasons that we derive from ethics, *why* we choose one method over another. Why, for example, we should not run an experiment in some situations. It is interesting to note that even if we find general agreement on at least many moral codes, how to do things in practice, we might often find disagreement in the reasons for why we perform —why we want to endorse certain judgments, or why we would perform certain actions and prohibit performing others. I will discuss this more. But first, let us probe our own intuitions.

We all have intuitions about what is morally right and wrong. What I want to do now is for us to investigate to what extent these intuitions might divert. For that purpose, let us discuss a few stories. First, imagine you are running a lab on vehicle security. You are on the cutting edge. You are the lab who can do the newest safety devices for road vehicles. And right now, you are developing such a device for lorries and you feel confident that you will be able to develop it. You also know that when you have developed that, then you will save 10 people in the first year of its implementation. But now you learn that there is another system for another type of car, for passenger cars, not lorries, that you could develop instead. Note however that you could not develop both technologies at the same

time. If you realize that, and you have the same chances of realizing that and implementing it in the same time frame, then you would save a hundred people. The question to you is, “Which of the two technologies would you choose?” Try to give an answer before you continue!

Here is my take on it: One technology could save 10 lives, the other technology could save 100 lives. And there are no other differences in the realization of these two technologies, or at least, none that we know of, and we cannot realize both technologies at the same time. We are forced to make a choice. What is the underlying principle here? Well, the choices differ with respect to their *consequences*, namely how many people are saved. Therefore, I would choose the technology that saves most lives.

Now consider another story. You are no longer an engineer, but a chief surgeon at a hospital. You have 100 patients who are dying from different organ failures in one year. And then there are 10 healthy patients, maybe they are just coming for a checkup. You could take the healthy patients’ organs, without them having given you their consent, which would imply their death, and then transplant these organs to the ill patients. Let us also say that the various organ failures are divided in such a way that actually, you can serve all the hundred patients from dying just with the organs from ten healthy patients. Now what would you do? Would you transplant or would you not transplant? Again, try to give an answer before you continue!

Let us begin by applying the same kind of reasoning as we did in the previous case. If we do not transplant, the consequence is that 10 people live, a hundred die. If we transplant, 100 people are saved, 10 people die. Clearly, we cannot have it both ways. Seen in this way, the result it is the same result as in the previous case. If we apply the principle that saving more lives is better than saving fewer lives, that leads us to accept transplantation. But perhaps something is missing in our reasoning. There is a difference between this scenario and the previous one. Here, as a surgeon you are directly involved in people's death, you are directly *causing* their death. This was not the case for the lorry manufacturer. The end result is the same in both scenarios, but in the lorry case, no one is directly intervening to actively cause death. If we believe it is wrong to directly cause someone to die, that will lead us to not perform the transplantations.

Interestingly, we have now shifted from talking about consequences to talking about behavior, or actions. This is an important difference that we need to address. The principle that leads us to the conclusion of not transplanting is not the one focusing on consequences, but the one that focuses on the particular properties of an action itself.

It seems that we have different ethical intuitions for different cases. Sometimes, we think behaving morally is justified by considerations of the consequences of these different actions. But in other cases, it seems that consequences are not enough. We might think that one also has to satisfy certain rules, or our actions need to satisfy certain rules, or certain

duties, such as do not cause anybody's, or do not directly cause anybody's death. Finally, a third intuition might have it that to behave morally is to exemplify relevant character traits, or virtues. So far, these are only intuitions. In the next part, I will discuss the theoretical frameworks that spell out these intuitions.

11.2 Three Frameworks of Normative Ethics

I have tried to show that we have different intuitions as to what the reasons for endorsing certain moral codes or particular moral judgments are. I want now to discuss three frameworks that explore these different ways of how to give reasons for endorsing moral codes. The first framework is consequentialism. Consequentialists argue that choices are to be morally assessed based only on their consequences. That view contrasts with the view of deontology, that to choose morally is to fulfill relevant rules or duties. A third framework, virtue ethics, argues that morality consists in having and exemplifying good character traits.

Consequentialism is perhaps the simplest framework. For the consequentialist, choices are morally assessed only by their consequences. This raises several questions. What kind of consequences are we talking about? How are they to be assessed? How are they to be evaluated? The most prominent example of consequentialism is **utilitarianism**. Utilitarians argue that we should assess the consequences of an action by the happiness or desire satisfaction (note that these two are not necessarily the same – there are many versions of utilitarianism) that the action brings about. If I consequentially evaluate this course, then I should ask each of you how much it satisfies your desires or your expectations or your evaluations of this course and compare that with alternative ways how to design this course. Individual evaluations need to be taken into account, and then be aggregated. The standard utilitarian view is that each individual counts the same, has the same worth. If we have a hundred students, we records their 100 evaluations, aggregate them, and then we compare it with alternative ways of how to run this course.

How is all this relevant for science and for evaluation of moral judgments in science? Well, consequences of choices in science are many. But perhaps we can categorize them in the following way. Particular choices might have consequences for the *epistemic goals* of science. Many of those fall into the standard epistemic methodology category, but various forms of deception also have moral relevance. We will come back to that when we talk about scientific misconduct. There are also consequences for the *institution of science*. Are scientists impartial? Are they seen as impartial? Are scientists trusted? Perhaps scientists now have a certain amount of trust that the public at large shows to them, but they might lose that if certain behaviors became more prominent. There are also the consequences for the *uses of scientific results*, as in technological progress, as in the creation of new risks or in the support of new technologies. So, consequentialism presents one way of how we can assess questions of morality in science.

However, there are a number of important problems that consequentialism generally is exposed to. For one consequentialism is extremely demanding. Every action has some consequence. But if moral judgment consists in the evaluation and comparison of different consequences of different actions, then there is pretty much nothing that is not morally assessed. If that is so, then morality is pervasive, encompassing every possible choice in your life. For every choice of action, you would have to take into account not just the consequences for yourself but for everybody else also. Seen this way, consequentialism rests on a very strong claim. The second problem facing consequentialism is that it might occasionally permit intuitively wrongful acts. For example, a consequentialist, at least a non-sophisticated one, would have said that you should transplant the organs from the healthy people to the ill ones, in the case discussed above, because you are saving more people by doing so. Consequentialists have developed further sophistications in order to counteract such non-intuitive results, but it remains a fact that in the particular case that we discussed, the consequences seem to clearly indicate that transplanting is better than not transplanting, despite many intuitions to the contrary.

Deontology: An action is permissible if it fulfils certain roles or duties. A duty might be to respect the rights of others.

Now consider an alternative framework, **deontology**. Deontology says that to choose morally is to fulfill relevant duties. Such a principle would seem to guide the negative choice of not transplanting, discussed in the transplantation case. Deontology, the name comes from the Greek "deon," which means duty, is the study of duties. The intuition of the deontologist is that some choices cannot be justified by, or only by, their effects or consequences. Instead, no matter how good or bad some consequences are, certain choices are prohibited, mandated, or required, in themselves. Everybody is familiar with such mandates or prohibitions. For example, in most countries, people have proper rights. Whether the division of property according to these rights is the most effective one, or the one that minimizes poverty, or maximizes economic development, does not matter. Thus, the consequences of such property rights are considered less relevant than the duty not to violate property rights. That is an example of a deontological consideration. Or we might say there are certain intentions that must never be permitted, like, for example, the intention to harm or kill someone. And in the transplant case, this rule seemed to lie behind the intuition of many of you.

Deontology is not just a defense of a list of duties that we have. There is a much more systematic approach behind it. It is the argument that we have rational insight into the set of rules and duties that we must adhere to. Many of these arguments are quite complicated, but let us consider one argument given by perhaps the most famous representative of deontology, Immanuel Kant, an 18th-century German philosopher who argued for **the categorical imperative**. This imperative is a meta-maxim. It is a principle from which all duties can be, or should be, derived. Kant's argument was that "if you have a maxim that you think should be a universal law, then you should act according to that maxim". It is an *imperative*, since it says what you should do, and *categorical* since it is absolute and unconditional.

The categorical imperative: An action is right if it is in accord with a generally defensible principle.

It is worth savoring the formulation of the imperative for a moment, because it is actually very sophisticated. The idea is not just to "do upon others as you want to be done upon yourself". Rather, it says that you should behave in a way such that your behavior would accord with a generally defensible principle. Consider a criminal that has been sentences to and put in prison. The criminal might be of the opinion that he is being treated too harshly by the law for his criminal actions. Maybe he even cries out that he wishes that nobody else would ever be put into prison. But if he was asked to consider what a society that do not punish those who violate the law would look like, he might agree that it is a good idea that we have a general principle according to which transgressors are lawfully punished.

Kant argued that all more specific roles and duties that we should adhere to according to deontology must be derived from the categorical imperative. Here are some examples that are relevant for scientists:

- We should treat humans as autonomous ends, as ends in themselves, rather than as instruments, as means for some other purpose that is not related to these individuals' own end.
- We should not intend to cause harm, irrespective of other benefits.
- We should speak truthfully, and we should defend the truth.
- We should respect intellectual as well as other property rights.

These are some examples of rules that I believe could be defended with reference to the categorical imperative that. The rules, in turn, guided some principles written in the Helsinki Declaration, for example, or that might be directly applicable when choosing between methods.

Deontology has some important advantages over consequentialism. In the first place, it makes life much easier for the scientist who wants to determine whether a particular research project or experimental design is in accordance with ethical rules. And this is something that you need to do now if you apply for new research grant that proposes, for example, an experiment with humans or with animals. In your application, you will need to account for in what ways you have considered ethical questions. The deontologist can go through a checklist and, in light of rules that are justified by the categorical imperative, determine whether their research is violating any of these rules. If not, then the research is ethically justified. For the consequentialist, things are not as straightforward, because the consequentialist needs to consider all possible consequences. And those are not as clearly codified as such a list would be.

Another advantage that deontology holds over consequentialism is that, because there is a finite number of principles that the deontologist believes one must satisfy, there are also areas in your life, certain choices that do not raise moral questions. This also seems to accord with many peoples' intuitions. Not everything is a moral question. There seem to be areas where we just do not worry about morality. A particularly important advantage of this is that this leaves some space for an agent's own concerns. Some decisions concern just one's own projects – for example

one's choice of partner, of career, of place to live. Everybody wants to realize themselves in some way. And as long as you are not violating a specified list some moral rules or duties, realizing these personal projects is not subject to moral considerations, according to deontology. For the consequentialist, the matter is not so clear. The consequentialist must contrast the effects of realizing their personal goals with the effects of alternative possible ways of acting or leading one's life. So, the demandingness, or over-demandingness, required by the consequentialist is somewhat curbed by the deontologist.

But deontology also suffers from a number of problems. First of all, if you only focus on rules, you might end up favoring actions with disastrous consequences. In the transplant case, we are basically saying that 100 people have to die because we cannot allow ourselves to sacrifice 10 people. But imagine a more extreme situation where we would have to actively kill one person in order to save one million people. The principle applies to exactly the same extent as in the case of weighing 10 versus 100 lives; we are still prohibited from intending to or directly causing someone's death. But at some point, the damage that we accept because we insist on this principle might be enormous. I am not saying that the principle is therefore wrong, but it shows that it would in fact be very difficult to actually implement such a principle.

The second problem is more practical but nevertheless very important. Because all rights and duties are categorical, there is no room for degrees. You cannot have a right to some extent, or you cannot have a duty 50% of the time. You simply have a duty that you have to satisfy. But if that is the case, then it is, first of all, difficult to incorporate **risk**. Many decisions in science are morally problematic precisely because they carry a certain risk – i.e. *uncertain* prospects of some harm. This uncertainty distinguishes most of the actual cases from the vignettes that philosophy teachers tell in seminars. Rarely do scientist or engineers decide to cause the death of 10 people with certainty. But they often have to make decisions that involve the risk of some people dying with *some* (presumably small) probability. But as the probably gets smaller, presumably, the strength of the deontological principle starts to weaken. But by how much? And how should one even express such a weakening of categorical duty? Deontologists have difficulties with incorporating risks in this way.

Risk: There are several definitions of risk. One is: the probability of an unwanted event which may or may not occur. See the literature for the "Risk assessment" lecture for more information.

Another issue with deontology is that weighing duties against each other when they conflict is not a straightforward endeavor. The consequentialist does not have a similar problem, because ultimately, we are trying to evaluate all the consequences with the same kind of currency, be it utility, happiness, or, in simpler cases of cost-benefit analysis, money. But the foundation of the deontology is that you have to respect the mandated duties. You have to respect one duty, but that might mean violating another duty. How do you trade this off? There are various answers to this question, but they often only work under very particular conditions.

The third framework that we will touch upon is **virtue ethics**. The idea behind virtue ethics is that morality does not consist in fulfilling any particular duties or bringing about the best consequences by acting in certain ways. Rather, for the virtue ethicist, morality consists in exemplifying good character traits. This idea goes back to the Greek philosopher Aristotle. Virtue ethics focuses primarily on persons, not choices or particular actions. It does not deny the importance of duties or of consequences, but it derives the evaluation of such duties and such consequences from the consideration of character of individual people. Virtues consists in—and this is something that is quite different from the other two accounts—an insight into human nature. Humans, Aristotle argued, have particular capacities, and they can acquire functions. Humans can, for example, reason rationally. They can also feel compassion. They can also feel the need to help someone. Some of these, Aristotle argued, are character traits that set humans apart from animals. And it is this ability to perform these functions that make a good character. In practice, this often means that we learn from others who we set as an example of moral character. You learn from that person by emulating them in a similar way as you might learn a particular craft or you might learn playing an instrument. The idea of the supervisor as a model of academic behavior (not just someone who helps you writing your papers!) belongs into this framework, too.

Virtue ethics: Morality consists in exemplifying good, or virtuous, character traits.

There is a lot of disagreement as to what the relevant virtues actually are. Examples that might come to mind include courage, respectfulness, resoluteness, sincerity, humility, reflexity; those are things that have been discussed in the literature. The advantage of virtue ethics over consequentialism is, similar to the deontological account, that it avoids over-demandingness. It only requires you to satisfy certain character traits. That does not mean that you have to consider the moral properties of every action. Interestingly, it also brings out a problematic feature contrasting it to the deontological account. The deontologist has a hard time explaining what motivates people to act morally, why they are willing to accept duties and principles.

It is much clearer from the virtue ethical account why a person would want to be a moral person. The answer is that being a moral person is a way of fulfilling your human nature. It is something that you would want to satisfy in a similar way as you might derive satisfaction from playing a musical instrument or being good in a particular athletic sport, or knowing something about a particular scientific discipline. However, virtue ethics also encounters a number of problems. In particular, the framework does not give clear guidance how to behave in particular situations. After all, we are now talking about character traits, and how these character traits play out for individual choices is often unclear. It is even more difficult to resolve moral conflicts between different courses of action with an account of virtue.

To conclude, I have discussed three frameworks of ethics that capture different parts of our moral intuitions, but fail in capturing others. In the

next section, I will begin applying these frameworks to specific questions of research ethics.

11.3 Morality and Experimental Design

Scientific experiments, especially those involving human or animal subjects, are subject to moral judgments, and must be constrained by moral codes of conduct. That was the clear message that emerged from the horrors of German concentration camps and Japanese warfare research, in response to which the *Nuremberg Code* of legitimate medical experiments was developed. But what were the ethical reasons for developing and justifying these codes of moral conduct? To see how the above frameworks can help answer this question, let us look at some examples of morally problematic experiments.

My first case, called the "Tuskegee Study of Untreated Syphilis in the Negro Male", bears its immorality already in the title. In the first place, it presents a racist perspective. Secondly, it studies untreated syphilis at a time when syphilis could be treated. This study ran from the 1930s to the 1970s, observing the health status of subjects who suffered from syphilis. At first sight, this might be considered an observational study, as an intervention seems to be missing. But upon reflection one realizes that the intervention was to *withhold* treatment at a time where all other patients diagnosed with syphilis would be treated (in fact, the US government at the same time ran expensive campaigns to encourage people to get tested for syphilis so that they could be treated).

Specifically, the subjects were deceived with regards to the length of the trial. They were initially told it was six months, and then it lasted for many decades. They all had syphilis, and their physical state was recorded regularly but they were not informed about the true nature of disease they had. Instead, some vague notion that they had "bad blood" was used. They were not informed about potential cures, even though those cures were in place at the time. In fact, those cures were withheld.

Now, this case is more interesting for a discussion of research ethics than, for example, the human experiments performed in Nazi concentration camps. Because, unlike the people being experimented on in the concentration camps, the people participating in the Tuskegee study were not prisoners. In some sense, they chose to participate in the study. The study, moreover, was performed in a democratic country, even though a democratic country that dealt with deeply entrenched racism. And it was that racism, that segregation in Alabama, that segregated, for example, the education system, that segregated the health system, which, in turn made it possible for this study to even work.

Hopefully it would not be possible to repeat this today. But at that time, it worked. These people had no access to proper schooling. Therefore, they lacked the ability to acquire information independently. And they lacked the ability, perhaps, to ask, or maybe the feeling or the self-confidence to ask, what do you mean by "bad blood"? Is there not something that can be done? The participants in the study were offered

certain benefits, such as free access to (the pretense of) health care, as well as, ironically and tragically, free burial. The participants chose to join the study, but today we would say that they did in fact not give their as **informed consent**. And this is the main reason why this experiment is clearly unethical.

Informed consent: a prospective participant is (i) informed, (ii) participating voluntary, and (iii) decision-capacitated.

Informed consent consists of three main parts such that a prospective participant is (i) informed, (ii) participating voluntary, and (iii) is decision-capacitated. Consent is *informed* if all the information relevant for the experiment is given—how it is performed, what the goal of the experiment is, what kind of side effects are known, and so on. Clearly, the Tuskegee study does not satisfy this. The researchers withheld information, with the particular reason that they did not want these people to then go and seek treatment, which would have made them drop out of this study.

Furthermore, the consent is *voluntary* if participation is not forced. The people in the Tuskegee study were not, strictly speaking, forced to participate, they were not prisoners. However, when considering voluntariness, we should also take into account social or economic pressures. People who are in a very bad or pressured situation—with no access to health care, suffering from a debilitating disease— cannot be expected to make a voluntary choice of participation, if they in fact are desperate to be relieved from these pressures and are willing to do almost anything for it.

This is why one should be suspicious when, for example, drug companies go to do test trials in less developed countries, because there the numeration, the payment that they can they give to people, can be enormously high in comparison to the salary levels in those countries. If you are at the subsistence minimum and you have a family, and someone is offering you a weekly or even a monthly wage in exchange for participating in the drug experiment, will you then reflect on what this might mean for your health? Perhaps not. In that case, voluntariness is not satisfied, as it was not in the Tuskegee study.

The third part consists in that the consent must be decisionally-capacitated. This requires that we do not merely hand information over and then let people fend for their own. Rather, we have to ensure that they are actually capable of processing the information we have provided them with. Now in the case of the Tuskegee study, it is not clear that the participants were capable of processing the information they received. One could suspect that these people were deprived of the relevant education, which would have helped them to understand what information was given to them (and which was not), and which would have also helped them to question the motives of the researchers who led the study. In other cases, such as if you aim to experiment with children, or you aim to experiment with mentally handicapped people, then you have to make very sure that you have satisfied this decisional capacitation among the subjects. In many cases, this is very difficult to

do, and in those cases, you might therefore not be able to perform an ethical experiment, because you cannot satisfy a proper capacitation for such subjects.

Let us now turn to the three ethical frameworks. Why is it important to obtain informed consent from participating subjects? From a consequentialist perspective, it is important to see that by ensuring their informed consent, we are enabling people to choose what is best for themselves. Individuals have access to private information (their desires and values, as well as tacit knowledge about their abilities and failings, their direct environment, etc.) that other people, even those in charge of helping them, rarely have. It would therefore be difficult for those outsiders to make decisions for these individuals. However, when making decisions themselves, individuals might still make mistakes; for example, they might lack abilities to organize this information, they might not understand what is at stake, or they might not have access to information that does not belong to this private sphere. Ensuring informed consent tries to prevent these mistakes from occurring, so that individuals have the optimal chance of making the best decision for themselves.

From a deontological perspective, the argument looks quite different. The deontologist requires informed consent in order to satisfy the duties of not lying, and, in particular, of treating others as autonomous beings. “Treat people as autonomous ends, not as means”, as Kant said. That principle was forcefully violated in the Tuskegee study. The researchers were using people as means, in order to find out more about syphilis. That information, of course, would never benefit the participants suffering from syphilis, and this was part of the setup.

From a virtue ethicist’s perspective, informed consent might in the first place be an exemplification of sincerity in the sense that you are being truthful and you are showing that you are willing to reveal all that you think is relevant. You also question whether your judgment is necessarily the best. Therefore, you want others to at least participate in making that judgment. The idea is that reflecting on your own limitations leads you to wanting to include others in making a decision.

We see how different frameworks would, in some cases at least, give different answers to how important particular implementations of informed consent are. Of course, the Tuskegee study is very clear in terms of moral judgments, but in real life, choosing how to design an experiment is not always so very clear. And that is where the interesting moral conflicts arises, where you will need to think about why informed consent is required. What are the reasons? That is when we need to consider the ethical frameworks, in order to discover these reasons.

One example, that I believe is a much more complicated, unclear case, is the that some German car makers, in particular VW, were recently associated with certain experiments involving humans, and were quite

severely bashed in the press for having performed or having supported these experiments.

Let us briefly consider what we know of these experiments. One experiment was performed at the Technical University in Aachen, Germany. Consenting subjects, students, were exposed to particular pollutants, in particular nitrogen dioxide, in a laboratory for a certain amount of time. Note that these exposure rates were lower than what was considered to be standard daily safe exposure rates until 2002. The experiment was run in 2018, so then it was not deemed safe – but you see that it is not *that* far off. Nobody expected that anybody was to topple over and die on the spot. The levels were such that could potentially lead to rashes, or maybe some increase in inflammation. But people were not expected to feel unwell. Furthermore, this study is the same as an earlier study that was performed, which found that in fact there are effects of such exposure rates. And it is funded by a group that is set up by the German car industry. The money comes more or less directly from carmakers, who, as you might know, were very much under duress in 2018, because they were trying to defend their continued production of diesel cars that are particularly associated with high nitrogen dioxide rates. And this experiment in the end found that there were no detectable effects of exposure.

The question then is, was this experiment morally permissible? The first thing to note is that the participants are consenting adult subjects. They are students, thus well educated. They were given information about this earlier study, as far as I know. So, the deontologist should be quite happy that the informed consent-rule is satisfied.

Furthermore, the associated risks are not excessive, judging from the information that we have here to consider. Given that informed consent was satisfied, subjects seemed to have willingly taken on the risk that the study included. Was it worthwhile for them to take the risk? The study in this form had been done before, but as I discussed in another chapter, it is an important part of science to check the replicability of experiments – replication tells us about the reliability of a result. So a consequentialist might conclude that the overall consequences of the experiment were positive: a small risk exposure of people who consented to it, offset by the larger benefit for society to get better evidence for claims about an important question in today's car-focused lives.

The most interesting aspect comes out when considering the involvement of the car industry. Surely, as researchers we are taking money from grant givers for particular research purposes. But in this case, I would argue that we are dealing with an integrity problem. Were the researchers really sincere in testing the results of such an exposure? Or because of the financing and the particular timing of the experiment at this time, do they perhaps exemplify a vice; a willingness to relent to temptations from the car industry in order to get some extra publication going, rather than for the real purpose of investigating these questions impartially. Here I think that, interestingly, the virtue ethicist has an edge over the other

frameworks, by being able to identify a problematic feature that the other approaches did not identify.

11.4 Scientific Misconduct

When producing empirical data, when developing ideas, or when testing hypotheses, scientists do not always make the right choices. This might be an issue of competence; sometimes scientists unintentionally fail to control an important background factor in an experiment, or unintentionally employ an inappropriate statistical tool to make an inference. Such failures of reasoning or lack of relevant knowledge might lead one to question the competence of such scientists, but the failures do not typically lead to moral blame.

It is an entirely different issue when scientists *intentionally* produce bad quality data or employ invalid inference rules in order to pretend that certain claims, important for their work, are justified. Take, for example, the case of Jan-Hendrik Schön, a German physicist working at Bell Labs in New Jersey from 1997 to 2002. During this time, he published a series of apparent breakthroughs with semiconductors, including 15 articles in the peer-reviewed journals *Nature* and *Science*. Schön became quite prominent internationally and received a number of prestigious awards. However, no other science team was able to reproduce Schön's experimental results. A subsequent investigation revealed that Schön refused to share his raw data that he presented different experiments by the same process data, and that some of his graphs, which purportedly had been plotted from experimental data, had instead been produced using mathematical functions. Consequently, Schön lost his job, many of his published articles were withdrawn by the scientific journals, and he eventually even lost his PhD title and his scientific awards.

There would seem to be a general agreement that Schön was to be morally blamed. Why is that? The ethical frameworks that we have discussed provide different answers to this question. Consequentialists might argue that Schön's actions are immoral because they undermine the public trust in science. His work passed all the peer reviewers and award committees, and they failed to discover that it was fabricated. Furthermore, his fraud also wasted a lot of resources. Other scientists trying to emulate his success started working in his field, only to discover that they could not achieve the same or similar results. Deontologists, in contrast, might stress Schön's fraudulent intentions; his wanting to deceive people and to fraudulently acquire scientific distinction are the basis of what makes his actions morally wrong. Virtue ethicists might instead stress the questionable character traits that his actions reveal. Schön's apparent desire to be celebrated for purported achievements that he knew where fraudulent seemed to exhibit a person who lacks sincerity and humility.

Schön fabricated data: he generated numbers and presented them as data, although they were not the result of any observation. But scientific misconduct is of course not restricted to **fabrication**. It also includes **falsification**. With falsification, genuine data is manipulated so as to, for

Fabrication: Intentionally making up data or results without scientific support to mislead the reader.

Falsification: Intentionally changing data or results without scientific support to mislead the reader.

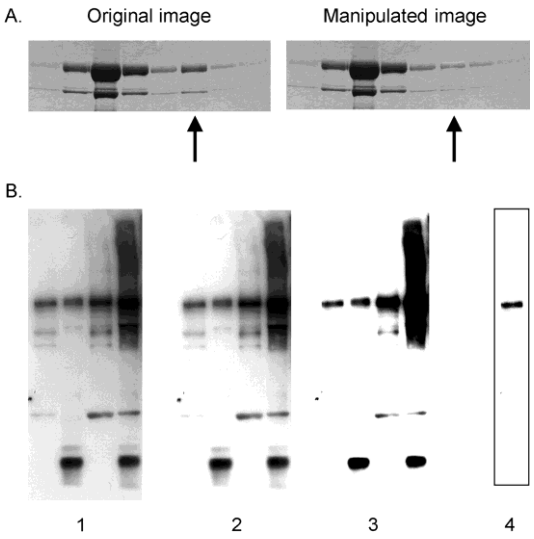
example, become evidence for certain claims, or inference tools are inappropriately used to suggest that data is evidence for such claims. Finally, **plagiarism**, the appropriation of another person's ideas, processes, results, or words without giving appropriate credit, is also considered a form of scientific misconduct. While the Schön case seems morally very unambiguous, many other cases of potential scientific misconduct are not. The first reason for this is that it is often difficult to find out whether *intentional* fabrication, falsification, of plagiarism has occurred. Scientific research often is a black box. It is very difficult for outsiders to gain insight into how a researcher arrived at her results.

Plagiarism: The appropriation of another person's ideas, processes, results, or words without giving appropriate credit.

For this reason, peer review is a very important institution in science—where other experts evaluating their peer's research. These experts must be competent and impartial to do this job properly. Reviewers typically offer their services for free, which might be an indication that many researchers consider it a moral duty to participate in peer reviewing. That they must be impartial also means that they should not gain from accepting a paper. Predatory journals threaten this impartiality. Predatory journals make money by accepting as many articles as possible and, therefore, pressure reviewers (if they have reviewers at all) to accept every submission. Unfortunately, the number of predatory journals has risen sharply over the last years.

But even when the peer review system operates properly, it does not always uncover misconduct, as the Schön case shows. Only the continuing scrutiny of other scientists, who are not reviewers, eventually led to the discovery of his fraud. Schön succeeded in avoiding this scrutiny for some time because he withheld his raw data. The Schön case shows how important **transparency** is to science. Had the raw data on which Schön based his claims been transparent to other scientists, his fraud probably would have been detected much earlier. A second reason for the difficulty of some misconduct cases is that good conduct is not separated from bad conduct by a clear line, but is differentiated on a continuum of ever-darkening gray areas. Most scientists presumably do not blatantly fabricate their data. But all scientists manipulate their data. They clean it, for example; they remove outliers, they construct a data model. This is perfectly acceptable practice, both methodologically and morally, as long as the process is transparent and scientists give good reasons for it on a case-by-case basis. But there is the risk of transgressing the point after which manipulating data turns into falsifying the data in unethical ways.

Scientific transparency: Freely sharing the material from which the conclusions were drawn.



Photographs, which are often used as scientific evidence, are particularly challenging. Take this shot of a western blot. It is somewhat dark, so an author might want to brighten it with Photoshop for publication. That is widely considered to be permissible, as the information contained in the image is not altered. But now consider this manipulation, where the darkness of a single band is reduced, indicated by the error. This is in fact a case of fabrication and, hence, constitutes misconduct because it violates the guideline that no specific feature within an image may be enhanced, obscured, moved, removed, or introduced. Detecting misconduct and distinguishing it from permissible practices is often not easy. But this does not excuse us from moral blame if we commit misconduct. While drastic falsifications like the Schön case might be rare, questionable research practices are all too common. A recent literature review found that almost 10 % of researchers, when directly asked, admitted to "modifying" data. Such practices include collecting more data after seeing whether results were significant, rounding down p-values, or selectively reporting studies that worked. Almost a quarter reported knowing of colleagues who had engaged in such questionable research practices. This raises the question of what can or must be done about this. Many agents have a responsibility to counteract such misconducts, including co-authors, research institutions, journals, and colleagues.

From a student perspective, perhaps the collegial dimension is the most important one. What should one do when learning of a colleague's misconduct? This brings us to **whistleblowing**. A whistleblower, in a research ethics context, is a colleague who exposes scientific misconduct. Such actions might have serious consequences for the whistleblower. She might be ostracized by her colleagues or fired from her job. So, should she do it? Might she even have a duty to do it? Here, our earlier considerations about the reasons for judging scientific misconduct again give different answers. The consequentialist can acknowledge the bad consequences of misconduct. But if she judges these consequences as not too big—let us say she considers the falsifying scientist to be just a small fish unlikely to make big waves, but she considers the consequences for herself to be really bad, then she will not consider it her moral obligation to expose the misdeed. A deontologist, in contrast, cannot consider such a trade-off. If it is her duty to speak the truth when a lot is at stake, then she will consider herself morally obliged to expose the misconduct.

Whistleblowing: Exposing illegal or unethical behaviour within an organisation.

To conclude, fabrication, falsification, and plagiarism are not just methodologically bad but morally blameworthy. Nevertheless, such misconduct is all too often found among scientists.

11.5 Authorship

Being an **author** is of particular importance in science. Although scientific progress is driven by the collective work of many contributors, it is the individual authors who are rewarded the reward. Being an author, or co-author, establishes credit for one's intellectual efforts. To many, it provides a personal sense of achievement. Crucially, it also contributes to one's professional reputation. Many even think of author papers as a currency stored by services like the Web of Science or Google Scholar, with which one can acquire academic appointments, promotions, and research funding.

Author: A person is an author of an article if the person has made a substantial contribution to (1) the conception of the work, (2) drafting the work, as well as (3) is involved in the approval of the final version, and finally (4) agrees to be accountable to all aspects of the work.

Authorship not only establishes credit, but also responsibility. By putting my name on a publication, I certify that it is the result of my own work, and that I performed this work to the best of my knowledge and abilities. I thus honor the trust of my readers, who cannot open the black box of research to certify the integrity of my work themselves. Furthermore, I also take on the responsibility to publicly defend the content of my work if challenged by readers. This all seems pretty obvious when we think of the authorship of 19th century scientists and scholars. The typical scenario then was the single author who, in independent work, developed ideas, made observations, tested theories, and put these results down on paper. Whatever was true or useful of this output, she owned, and whatever was false or damaging was her responsibility. Today, scientists sometimes do not live up to this scenario, not least because science operates differently today. Nevertheless, it remains an important normative ideal that guides moral judgments today. In order to understand this ideal better, it is helpful to consider the definition of authorship that is almost universally accepted among scientists today.

This definition was first proposed by a group of medical journal editors in 1978. They stipulated that an author be a person who is involved in, first, a substantial contribution to the conception of the work and, secondly, the drafting of the work and, thirdly, giving final approval of the version to be published. Note that all these conditions must be satisfied for authorship. Just failing one of them disqualifies a person as an author. In 2003, the committee included a fourth requirement, which explicitly puts the responsibility for any misconduct on all co-authors. This is that the person agrees to be fully accountable for all aspects of the work, including those that she did not produce herself. Clearly, this definition specifies a normative ideal—what authorship should be, not what it always is. Therefore, we need to ask why authors should satisfy these criteria. What are the ethical reasons for them?

Unsurprisingly, views differ on these ethical reasons. Consequentialists typically argue that these criteria are necessary in order to maintain the impartiality of science, to preserve public trust in science, to uphold the incentives for individual scientists to work for personal achievement, and

to save resources. Deontologists, in contrast, might argue that the criteria prevent the violation of intellectual property rights and assign responsibilities to the right persons. Virtue ethicists, finally, might see these criteria as the consequences of virtuous character traits, like sincerity, humility, resoluteness, and reflexivity.

The normative ideal is, of course, only relevant when people do not always live up to it. Let us therefore look into some ways in which scientists violate the requirements of authorship, and what reasons we have to morally blame them for these transgressions. The first violation of the criteria is plagiarism. In Latin, *plagiarius* means kidnapper, and the plagiarizer does just that. He kidnaps someone else's text. More precisely, to plagiarize is the action or practice of taking someone else's work, idea, etc., and passing it off as one's own. It is literary theft. Of course, almost all scientists rely on other people's work and ideas. When they do so correctly, they clearly attribute those ideas through citations and references. A plagiarizer, in contrast, does not indicate their sources of these ideas, but passes them off as his own. This violates the first authorship criterion—the plagiarizer did not contribute substantially to the conceptualization of these passages, but publishes them in his name, thus pretending he did. From a deontological perspective, a plagiarizer is clearly morally blameworthy, for he violates intellectual property rights and lies to his audience. Likewise, from a virtue ethics perspective, as the actions indicate a lack of sincerity and humility. From a consequentialist perspective, things might be more ambiguous, at least for those plagiarizers who put up a kind of "nobody gets hurt" defense. "I am just taking from public domain material," some of them might say, "so I am not violating anybody's legal property rights." But this misses the point.

Even if the consequences for those stolen from are not particularly severe, the consequences for those pretended to typically are. Plagiarism seeks the unmerited increase to the plagiarizing author's reputation or the obtaining of academic credit achieved through false claims of authorship. Those who give the credit and honor the reputation are deceived, and thus harmed. A sincere consequentialist, by considering the consequences to the audience thus lied to, will be unambiguous in the moral blameworthiness of plagiarism. A particular kind of plagiarism is self-plagiarism, the use of one's own previous work in another context without citing that it was used previously. It does not violate any property rights, so it cannot be morally wrong for that reason. But as I have argued above, this is not the only reason for blaming the plagiarizer. Instead, the self-plagiarizer is blameworthy because she deceives her audience. She claims to present something novel while, in fact, it is not. Furthermore, she also wastes valuable resources. Other scientists will spend time reading through her self-plagiarizing contribution, only to find out that it does not contain anything new.

While plagiarizers illegitimately claim authorship to establish unearned credit, **ghost authors** hide authorship, perhaps to avoid responsibility or to conceal conflicts of interest. Unfortunately, such practices are more common than many might suspect. For example, Monsanto, the producer

Ghost author: A person who is not stated as an author but fulfils the criteria to be stated as an author.

of Roundup, one of the dominant brands of the controversial glyphosate herbicides, was discovered to have systematically approached scientists for ghost authorship. In one of the discovered emails, Monsanto executive William Heydens wrote in 2015, "we would be keeping the cost down by doing the writing, and they [the scientists] would just edit and sign their names, so to speak" (*National Public radio*, 2017-03-15). Heydens further wrote that this is how Monsanto had handled an earlier paper on glyphosate's safety. Here, the true author of the report, according to the normative criteria discussed above, would not appear as the author of the text, but other people would.

So, why is ghost authorship not permissible? Consider arguments from each of the three ethical frameworks. A ghost author writes or co-writes a manuscript, but is not listed as an author. This is morally not permissible for a number of reasons. For the consequentialist, hiding an author threatens the impartiality of science, as it conceals potential conflicts of interest, and it ultimately undermines public trust. For the deontologist, the ghost author shirks responsibility for what he or she wrote. For the virtue ethicist, the ghost author exhibits a lack of resoluteness, the virtue of staying with one's work, forging on despite difficulties, and reflexivity, the virtue of being critical enough of one's work, making due allowances for one's own bias.

I started out with the picture of the lonely author in his study. That is less and less a reality in science. Over the last decades, the number of multi-authored papers has risen dramatically. This has led to a number of specific moral issues regarding co-authorship. Who should be a co-author? This is the primary question. And the phenomena of presenting someone as co-authors, even though they do not meet their authorship criteria, are described as honorary or gift authorship.

Gift authorship: A person who is stated as an author, for instance for social reasons, but who do not fulfil the criteria of being stated as an author.

Gift authorship is even more prevalent than ghost authorship. In a 2011 survey, 14% of all scientists surveyed admitted to have been involved in cases of gift authorship, while only 7% admitted to having been involved in cases of ghost authorship. Typical but morally difficult examples here are the inclusion of grant winners who did not substantially contribute, or PhD supervisors who did not participate in drafting, or the exclusion of a group member only because he will be on the job market soon. The ethical reasons here are very much the same as in the plagiarism case. Gift authorship is morally wrong because it deceives the audience and, through such deception, incurs possible harms. The subtlety here lies more with determining what kind of contribution merits co-authorship. Remember that the first authorship criterion was substantial contributions to the conception of the work. Thus, the grant winner who, in the grant application, already had developed the study design might be considered a co-author if she also meets the draft and final approval conditions, while someone only obtaining funds would not. Similarly, the supervisor who contributes to the data analysis decision or performs some interpretations herself might qualify, while the supervisor who merely comments on the paper does not. More generally, those who merely run an experiment

provide technical support, collect data, proofread or comment will not qualify as co-authors, although, of course, they should be thanked in the acknowledgments.

Another tricky issue is to determine the order in which co-authors are named. A standard view is that the author who has made the most substantial contribution should be credited to be the first author, but this requires measuring and weighing the different co-authors' contributions, which is not an easy feat. In any way, such a determination clearly is a dynamic process, and should never be based on a pre-determined decision. It also should include *all* individuals involved in the study!

An interesting alternative is to list the exact contributions of each co-author in a footnote of the published paper. This makes it more difficult for individuals to claim rewards not belonging to them, and reduces the importance of order, as authors receive credit according to the personal contribution. It also makes it easier for other scientists to contact the right person when they are seeking further information. To conclude, who is presented as an author of a scientific paper and who is not raises multiple moral issues relating to harmful consequences to the readership, as well as to scientific institutions, to property rights and norms of responsibility, and to virtues of sincerity, humility, and reflexivity. Reference to the standard normative ideal of authorship will be enough to clarify many cases, but sometimes negotiations involving all contributors are required.

11.6 Moral Career Choice

When we discuss questions of whether to choose a particular experimental design, whether that design is morally acceptable or morally permissible, we are basically discussing method choice. Ethical considerations and reasons provide additional reason for choosing one method, or a particular way of experimental design, over another. In this way, they expanded our previous method choice discussion to not only include scientific epistemic goals, but also ethical reasons.

But scientists also face moral questions about choosing their research projects or, more generally, choosing their careers – thus choices on a more general level than between specific methods. In the first place, some research projects carry risks, including environmental dangers, health risks, or economic, social, or cultural harms. Should a scientist not pursue such projects?

The simplest case is when the project aims to produce only harm. Think of Dr. Strangelove, the antihero played by Peter Sellers in Stanley Kubrick’s film of the same name. Strangelove, an ex-Nazi scientist now advisor to the president dreams about building the doomsday machine that destroys all of humanity. Participating in such a project would clearly be morally blameworthy.

But perhaps real baddies only exist in the movies. Real scientists typically face projects where the main goal aims to generate benefits, but there are some risky side effects. One area where this plays an important role

concerns dual use technologies. Dual use is defined as those goods or products and technologies that are usually used for civilian purposes, but that can be easily repurposed for military purposes, in particular for illegitimate military purposes. In the first place, we might think this is a question of government regulation, but I think we can apply the dual use challenge also to the question of design. Should a scientist participate in designing technologies that can easily be repurposed to do some serious harm?

With respect to that question, I think the deontologists and virtue ethicists have quite a different answer than the consequentialists, because the deontologists in particular focus on individuals' intentions, or maybe on the direct causal link. According to the deontologist, you must not intend to harm, and you must not intend to deceive or to kill. But the design of the potential dual use technologies clearly does not intend the repurposing for some harmful purposes, so therefore, it is not morally prohibited to design such technologies. Now contrast that with the consequentialist perspective. From the consequentialist view, what is important is that we consider all possible consequences. So here, the designer cannot excuse herself by saying, look, I did not intend this for these consequences, and it just so happened that this turned out to be the perfect murder weapon. Instead, the consequentialist perspective would require for a designer to use available knowledge, consider such possible consequences, and base their decision on that. Thus, the consequentialist takes a much clearer and stronger perspective on dual use than the deontologist would.

Dangerous and risky technologies is not the only concern that arises when considering obligations or prohibitions of research topics. Another issue that arises concerns **frivolous research**. Maybe the best illustrations are those that are regularly awarded by the IgNoble for the least useful research publications. It is a wonderful list! I have randomly grabbed two that were recently awarded. One publication investigates the effect of different types of textile on sexual activity. The other one investigates the effect of mirror scratching on itch relief. That seems amusing, but surely there is some interesting question, and the research was published in serious journals in the respective disciplines.

Frivolous research: Research conducted (at least as far as can be determined) purely for the fun of it.

Now contrast the topic for these publications with what the World Economic Forum thinks are the biggest challenges of our time. We have a bunch of highly educated people who have spent many years at university. In Europe, this is supported with public money, and if the people behind the publications are still employed by a university, then they are likely to still receive public money. They are investigating these things that seem amusing, but perhaps little more than that. At the same time, people are dying due to lack of food security, and we are threatened with rising sea levels, and suffering from many social destabilizers that we do not really have answers to. Should these people morally have chosen to contribute to the solution of those problems instead of doing whatever they think might be interesting for their own diversion?

Consequentialists, unsurprisingly here, have a very clear answer. This is maybe best exemplified by a recent group that call themselves **effective altruists**. They are philosophers who have actually come up with a very concrete proposal. They argue that every human spends, in their lifetime, about 80,000 hours on work. You should think of that as your capital that you can use to do good or less good – even if not necessarily bad, you can still waste it. You can sit by the pool or you can investigate the effect of textile on sexual activity. But you could also make use of that time in order to educate and train yourself in the best possible way in order to address food security, or rising sea levels, or social destabilizers. And they argue that that is exactly what you should do.

Effective altruism: A movement aiming to benefit others through the most efficient methods possible.

Thus, there is no free choice of research topic or even of career path. There is a moral obligation to choose those careers and those topics that are most relevant. This does not mean that we all have to become doctors and cancer researchers. The argument has to still reflect the diversity of human abilities. Some people might be particularly adept at investing or of planning and organizing financial investment. Those people should go in those areas, irrespective of what their personal plans are, and they should make as much money as they can in order to then give that to causes that promote the improvement of welfare overall in the world. Other people, who are particularly capable in academic research, should put their investigative capacities to investigating the most pressing issues so that people can be helped. And here, the World Economic Forum is just an example. What is important is that for all of you, is that research choices, career choices are not just a matter of your private preferences. They are a matter of moral judgment.

There is a deontological counterargument, and I think this shows how far consequentialists and deontologists sometimes can come apart. When you pursue the 80,000 hours to the extreme, then you are making yourself an instrument of the betterment of the world. And according to Immanuel Kant, no one, including yourself, should be considered as a means to some other purpose. Rather, they should all be treated as autonomous ends in themselves. The moral code should not prevent you from autonomously deciding what you are, and how you want to develop. This applies to a very similar extent to virtue ethics too. There is a sense in which you individually should flourish in developing your virtues, and you can exemplify these virtues in different ways. The consequentialist is criticized, for example, by the 20th century philosopher Bernard Williams, for not allowing any partiality to one's own project. Instead, Williams argued that the consequentialist moral argument leads to an alienation of people from their own projects, and this might lead to a distancing from moral theory.

To conclude, I have presented you with reasons for endorsing moral claims. Moral claims concerning methodological and research projects are plenty, and there are many moral codes for scientists. What I have focused on in particular however, is the normatively valid justifications for such moral claims. To this end, I presented three frameworks, none

of which captures all of our moral intuitions. That was also not the purpose of this lecture. Rather, it was to provide you with some concepts that can help you expand and improve your own reflections that you can then use to analyze specific problems and find your own solutions to them.

Part 12 – Definitions

It might not be obvious at first glance that the meaning of words or, to express it more formally, the notion of linguistic meaning needs to be the subject of philosophical theorizing. After all, in everyday life, language is for the most part not something we experience as in need of theoretical treatment and subsequent explanation. Language is like a program that runs in the background of our lives: it is a prerequisite for being able to perform a lot, if not most, of the tasks that we perform on a regular basis, but we rarely take notice of it – let alone experience it as problematic.

However, upon further reflection, we see that the study of meaning and definitions (so-called *semantics*) does belong in a course on scientific methodology. The meaning of our words determines what we are talking about, which in turn determines what it would take for our statements to be true (that is, what the world would have to be like). Now, science is a generally construed method of satisfying the "justification" criterion of the classical definition of knowledge (see part 1 if you do not remember what that is). Whether or not we are *justified* in accepting (or rejecting) a particular claim in light of some data depends, in part, on what the world must be like in order for the claim to be true, and thus on the meaning of the words in the claim. In short, the truth-conditions of a given claim are an output of the meanings of the words embedded into it. This is why scientific methodology cannot completely separate itself from semantics.

But why the focus on definitions? After all, unlike meanings, definitions are artificial constructions. When we grow up, we typically learn new words by repeatedly hearing others say them, and inferring their meanings from the contexts in which they are uttered, rather than memorizing their definitions. Over time, our minds build up a type of "mental lexicon" that we continue to refine all throughout our lives. Definitions, however, almost never come into play in this process.

Nevertheless, definitions can be useful for bringing to surface the question of what we mean by the words we are using. This is especially important when there are unclarities embedded into these words. And, as will be discussed more below, in scientific context it is often important to get rid of such unclarities to the furthest extent possible.

Definiendum – What is to be defined.

Definiens – What provides the definition

Definitions consist of two parts – a term to be defined (a **definiendum**) and a description intended to represent the meaning of that term (a **definiens**) – that are linked together by a connective that expresses either identity or logical equivalence. So, for example, we could define the word *bachelor* by saying that…

> *A bachelor* (definiendum) *is* (i.e. is identical to*)* *an unmarried, adult, male human being* (definiens)

Importantly, the very fact that something is a definition does not depend on how successful, or useful it is. The following is another definition of *bachelor*. It is just as much of a definition as the previous one, although it is obviously not as close to being an accurate representation of the meaning of the word *bachelor* in ordinary English.

A bachelor (definiendum) *is* (i.e. is identical to*) a married woman that speaks Russian* (definiens)

Now, definitions come in two flavours: **lexical definitions** and **stipulative definitions**. The difference between these is that while lexical definitions attempt to represent meanings that are assumed to already exist, stipulative definitions bring meanings into existence: they *make it so* that the definienda[5] have a particular meaning (namely the one represented by their associated definiens). A lexical definitions is a *hypothesis* about what the meaning of a given definiendum is, in a particular language or context ("what people usually mean"), while a stipulative definition is a *constitutive declaration* of a definiendum's meaning, in a particular language or context ("what I now intend the word to mean").

Lexical definitions: What is commonly meant by a word in a certain language or context.

Stipulative definition: What a word is intended to mean in a certain language or context.

Interpreted as lexical definitions, the definitions above are competing hypotheses about what the actual meaning of the word *bachelor* is (in English). As such, they can be evaluated with respect to how **accurate** they are. Obviously, the former is better representation of the meaning of *bachelor*, than the latter. However, even the former definition is arguably not perfect. For one thing, it might be **too wide** – it covers too much. Its definiens applies to some things that is not included in the reference class of its definiendum (as it is used in the targeted language or context). Consider for example that the pope fulfils all the conditions listed in the definiens (that is, the pope is an unmarried, adult, male human being); even though most of us have the intuition that would be incorrect to say that the pope is a bachelor. Being a bachelor seems to imply that you are planning on marrying in the future. Furthermore, it might be **too narrow**, in the sense that its definiens fails to apply to some things that are included in the reference class of the definiendum. In our example, such cases are less obvious, but perhaps it is imaginable that someone who strictly speaking is not a human being (but, say, a sentient, intelligent humanoid robot) can still count as being a bachelor.
Importantly, these properties – broadness and narrowness in relation to a target meaning – are not mutually exclusive. Defining a *bird* as a *flying animal*, for example, is both too broad and too narrow at the same time. It's too broad since the description *flying animal* is true of bats, even though bats are not birds. It is too narrow since penguins are not flying animals, even though penguins are birds.

Accuracy (definitions): To what extent a lexical definition captures the common meaning.

Wide: A lexical definition is too wide if its definiens applies to more things than it should.

Narrow: A lexical definition is too narrow if there are things that its definiens does not apply to, which it should.

The definiens of a definition consists of criteria that are intended to be (i) *individually necessary* and (ii) *jointly sufficient* for something being in the extension, or reference class, of the definiendum. For example, this means that the definition should be interpreted as saying that, (i) in order for something to be a bachelor, it must be a human being, it must be unmarried, and it must be an adult, and (ii) nothing else is required. Evaluating the accuracy of a lexical definition requires consulting one's own, and perhaps also other speakers' linguistic intuitions about whether the criteria *really are* individually necessary and jointly sufficient.

[5] This is the plural form of the word *definiendum.*

Interpreted as stipulative definitions, neither of the two definitions above is more accurate than the other. Stipulative definitions cannot be more or less accurate, because accuracy can only be estimated in relation to a target. Stipulative definitions do not have targets. Rather than trying to represent an already existing meaning as accurately as possible, they determine what the meaning of their associated definiendum is in a particular context. One typical reason for crafting a stipulative definition is that existing lexical definitions of a definiendum are defective (e.g. too broad or too narrow). For example, ordinary English usages of 'mass' and 'weight' seems to be synonymous, while physicists offer different stipulative definitions for each term. Another typical reason for crafting a stipulative definition is a new scientific discovery: for example, on 11 February 2020, the International Committee on Taxonomy of Viruses adopted the official name "severe acute respiratory syndrome coronavirus 2" (SARS-CoV-2) and gave it a stipulative definition. As research progresses, this definition has been, and presumably will further be, adjusted.

Unsurprisingly, the domain of application of a stipulative definition is typically a lot more narrow than that of a lexical definition: while an individual or organization might have the authority to determine what a particular term is going to mean in the context of, say, a scientific paper or a company-internal system of rules, no one can simply decide that a certain word is going to have a particular meaning in English in general, and thereby make it so.

Stipulative definitions can be better or worse. Their quality can be estimated on grounds of how well they facilitate the fulfilment of the purpose for which they were created. In this sense, stipulative definitions are **purpose-dependent**. It is not possible to say whether a given stipulative definition is any good or not, unless we know what it is supposed to be used for. However, in most cases, the purpose of a stipulative definition is to bring clarity to where there previously were unclarity. In many cases, a stipulative definition therefore will be a refinement of an already existing lexical definition that is found to be wanting in some way, and that can be improved by making it more precise. Therefore, it is safe to say that, most of the time, the more **precise** a stipulative definition is, the better it is.

Consider for example that the public transport organization in Stockholm, SL, defines an *adult* as *someone who is 20 years or older*. This is approximate to the lexical meaning of the word, but is more precise. The lexical meaning of the word *adult* is **vague** (or imprecise). This means that there's no clear-cut distinction between the things it applies to and the things it does not apply to: if we consult our own intuitions about what it means to become an adult, we realize that it is not something think of as happening overnight – let alone from one second to the next. Rather, becoming an adult is a gradual process. However, SL needs there to be a clear-cut line because they need it to be completely determinate whether someone is obligated to pay the full fee for a ticket or not (non-adults get discounts). Similarly, colour terms such as *red*, *blue* and *yellow* are paradigmatic examples of vague terms. Vagueness stems from the fact that our language is much more coarse-grained (or "chunky") than the world itself. Colour terms are thereby vague because, while they themselves constitute pretty clear-cut "chunks"[6], the actual sets of nuances they refer to are

Purpose dependent: The quality of a stipulative definition is dependent on its purpose and is therefore purpose-dependent.

Precision (definitions): To what extent a stipulative definition provides clarity in its context.

Vague: A word is vague if there is no clear-cut distinction between where the word can be applied and where it cannot.

[6] Arguably, this is true even though color terms are sometimes used as gradable adjectives: as in *blue, bluer, bluest.*

fuzzy. Where, for example, do we draw the line between green and yellow? In other words, where does the reference class, or extension, of the term *green* end, and that of the term *yellow* begin? Now, compare this with a term like *biological sibling*, which arguably is not vague. The line between the sets of individuals this term does and does not refer to is sharp. Either someone is or is not the biological sibling of the person under discussion – it's never (or at least very rarely) somewhere in between. Consequently, *biological sibling* is presumably not vague.[7]

A lot of the time, the purpose of constructing a stipulative definition is to reduce vagueness. In scientific contexts in particular, vagueness is considered a problem since it can make the empirical implications of hypotheses indeterminate. Consider for example the following hypothesis:

Red mushrooms are more poisonous than other mushrooms

Unless we "stipulate away" the vagueness of the term *red*, there is a wide variety of possible observations that neither constitute evidence in favour of the hypothesis, constitute evidence against the hypotheses, nor are completely irrelevant to the hypothesis. What are we to say, for example, about mushrooms whose colours are in-between paradigmatic red and paradigmatic purple? In summary, stipulative definitions are used in science in order to increase the precision of statements by reducing vagueness in order to make their truth-conditions more determinate.

Importantly, vagueness should be distinguished from **ambiguity**. This is when one and the same sound, or string of letters, has two or more distinct meanings. *Bridge*s, for example, can be found on violins, in songs and over rivers. The word *space* can refer to the distance between planets and stars as well as to the empty area in a room (and also, perhaps metaphorically, to the state of being alone – as in "*I need some space*"). These words are ambiguous. You can think of them as distinct words (with their own associated meanings) that just happen to be spelled and pronounced the same way – either by pure coincidence, or because they have a common history.[8] Ambiguity is therefore a phenomenon at the level of the word itself – that is: the sound or string of letters. Vagueness on the other hand is a phenomenon at the level of the meaning.

Ambiguous: A word is ambiguous if there are two distinct meanings of the words, in the way that they could be interpreted as two completely different words.

Now, consider the following example. We are to define *recklessness*. In other words, this is our definiendum. The context is that we are trying to create a framework to identify young people who are at risk of developing social problems, and we want to see if recklessness is one factor which could be identified as a criterion. I would start with consulting my own intuitions about the meaning of the term, I would consult dictionaries and read newspaper articles and scientific books where this term is used. Using these, I would try to find the commonalities of these usages and form this into the definiens of a

[7] Some linguists and language-oriented philosophers argue that most, if not all, so called lexical terms (noun phrases, verb phrases and adjective phrases) are vague – and, by implication, even the term *biological sibling*. Whether or not this is correct is up for debate. However, it is definitely true that vagueness is more common than one might think. Even artifact words such as *chair* and *sofa* could be argued to be vague, due to there not being a clear-cut limit to how wide a chair be before it becomes a sofa.

[8] Whereas words like *bridge* and *space* exemplify *lexical* or *semantic ambiguity*, there is also what is sometimes called *syntactic* or *structural ambiguity*. This is exemplified in sentences such as '*I saw the man with the binoculars*', which can be interpreted as saying either that I saw the man that had the binoculars, or that I used the binoculars to see the man.

lexical definition: perhaps “behaviour where one is unconcerned about risks and negative consequences”. I would then test this definition against my source material and see if there are there are cases indicating that my definition is too broad or too narrow – something is recklessness according to my definition even though it should not be, and so on. However, even with a lexical definition there might be cases which I intentionally choose to ignore, because not everyone uses words in the same way. I might have to disregard cases to have a simpler definition, more consistent with the general use. Now, taking this lexical definition, I move on to create a stipulative definition for my purposes. Then the definiens might be “in at least three reported cases have performed actions even though they meant low benefits and large risks”.

Of course, this definition includes “low benefits” and “large risk”, which then again might need to be defined in a new, separate process. While this could in principle go on for a very long time, the process is not infinite, since all words are defined using other words, and for most practical purposes it becomes clear when the remaining words are clear enough without needing a further definition.