3. Three ideals of science

In this chapter three ideals about scientific methods, procedures, objectives and what constitutes good science, which are assumed to be widespread in the public understanding of science, are explored in relation to climate science. However, I would like to make clear here that the aim is not to present some kind of sociological study that makes some general assessment of the public understanding of science and from there establishes and categorises these ideals. Instead, I will take a top-down approach to this question and will infer from a number of instances from the history of climate science – as established in the introduction to this book – where science sceptics were very effective in discrediting particular climate research in the eyes of the public, that they benefited from the general popularity of certain idealised assumptions about how science operates. The primary objective of this chapter is to show, based on the work of philosophers of science over the last century, why a failure of science in general and climate science more specifically to live up to these ideals is not a sign of inadequate science. Further, it will be analysed why the failure of these ideals becomes particularly visible in the context of climate science.

3.1 Value-free science

3.1.1 Introduction: values in science

Science as a value-free endeavour has long been and continues to be an ideal upheld by the public, scientists and (to a certain degree) philosophers alike as a definition of what constitutes good science. This is no different when it comes to climate science. As we have already seen in the introduction to this book, the accusation that climate scientists are biased and, thus, not objective in their research is at the core of many climate-change deniers' arguments.

The value-free ideal of science is often intertwined with a certain understanding of *scientific objectivity*, where objectivity is defined by the relationship between science and (social) values. This kind of objectivity may take several different forms. For a better understanding of how varied scientific objectivity can be interpreted in relation to values, it is worthwhile to take a closer look at this type of objectivity in Douglas's classification of scientific objectivity introduced in the last chapter (2004). Douglas argues that there are (at least) three versions of this kind of scientific objectivity. The first option to define objectivity in this way (that is, in relation to values) is to claim that values should not be allowed to override evidence. This, as Douglas calls it, *detached objectivity* is by comparison a rather broad definition. It does not completely rule out that there is an appropriate role for values in science. However, there is also a narrower, more common understanding of scientific objectivity which does exclude values in (almost[1]) any form. This *value-free* ideal is what I will primarily discuss in this chapter.

The third kind of value-related understanding of scientific objectivity which Douglas identifies is that of *value-neutral objectivity*, a view of science, which acknowledges values in science (to a certain degree), but scientists are urged to take a middle-ground position. This is a point of view on the role of values in science that will not be discussed here in more detail. Suffice it to say that taking no sides at all might be undesirable in certain situations, if what lies on one side of the value spectrum is otherwise considered absolutely unacceptable, such as racist or sexist positions.

In the following, I will discuss why the value-free ideal cannot be maintained in the case of climate science and science in general. To that end, I will first outline the historic background. A look back in history helps to better understand the value-free ideal in general and how it has risen to such prominence in the last century. Before actually turning to the debate about the role of value judgements in climate science, I will also take a closer look at the discussion of value judgements in the context of inductive-risk assessments, which has taken up a prominent place in philosophical debates about the role of values in climate science.

[1] Philosophers of science who advocate for this strict value-free ideal of science commonly acknowledge that there is a small number of "epistemic" values. These are considered to have an appropriate role in science compared to so-called "non-epistemic" values which are generally, according to this view of science, considered inappropriate; a distinction which will be further discussed in this Chapter.

3.1.1.1 The rise and fall of the value-free ideal

Historians and philosophers retrace the origins of the separation of values and science to Francis Bacon and the beginning of modern science and philosophy (Carrier, 2013; Douglas, 2009; Proctor, 1991). Bacon voices concern that moral deliberations would deceive men in their pursuit of scientific knowledge and prevent them from fully dedicating themselves to the advancement of science (Bacon, [1620] 1863). Another forerunner of the value-free ideal, which is often cited and should at least be mentioned here, is David Hume's distinction of "ought" and "is" and the notion that "ought" cannot be inferred from "is" (Hume, [1739–1740] 1888, p. 469).[2] However, it was not until the late 19th century and the rise of social science as a scientific discipline that the notion of value-free or value-neutral science, as we know it now, emerged (Proctor, 1991, p. 65). At the beginning of the 20th century the value-free ideal was prominently supported by the German sociologist Max Weber. Weber was the leading intellectual in the *Werturteilsstreit* advocating for a strict separation of science (specifically social science) from values. He was especially concerned about the university professor who might push their political ideals onto their 'defenceless' students and pass them on as scientific facts. Value judgements can also have an undesirable effect on science itself, according to Weber, to the extent "that whenever the man of science introduces his personal value judgment, a full understanding of the facts *ceases*" (1946, p. 146).

But even though the value-free ideal had a prominent and committed advocate in Weber, it did not prevail until the middle of the last century. As Douglas (2009, pp. 44–46) has shown, there was still a lively debate about the relevance and necessity of values in science and what form a value-free science

[2] Proctor (1991, p. 65) remarks that Hume was not the first to make this distinction. He also notes, referring to Hampshire (1949) and MacIntyre (1959), that Hume himself did not fully separate "ought" from "is" and only ever meant that the former could not be derived logically from the later. "But ideas live a life apart from the intent of their authors. In the mid-twentieth century, Hume's call for a separation between "ought" and "is" became a rallying cry for scientists and philosophers defending the neutrality of science" (Proctor, 1991, p. 61). Proctor also stresses that Bacon's and Hume's position here must be seen in the context of a wider move from philosophers and scientists to separate science from religion and questions of ethics as well as a new recognition of subjectivity in science, which is most visible in the distinction between primary and secondary qualities (Proctor, 1991, p. 54).

should or could even take throughout the first half of the 20th century. Prominent philosophers of that time such as Robert Merton and Ernst Nagel voiced at least mixed feelings about the topic.[3] Others even argued that values in science were unavoidable or even necessary.[4] Particularly strongly debated were value judgements in the context of inductive risks. Richard Rudner (1953) and C. West Churchman (1948) argued that under certain circumstances scientists cannot but must make value judgements in their scientific deliberations. They note that whether or not scientists accept or reject a hypothesis depends not just on the epistemic evidence but also on the severity of possible social and ethical consequences of a wrong decision (a longer discussion of the inductive-risk argument follows in Chapter 3.1.2). Popular counterarguments against this reasoning are that such value judgements should be handed over to the public, while the role of the scientists is only to ascribe probabilities to hypotheses (Jeffrey, 1956), or that all value judgements scientists have to make can be solely determined by inner-scientific "canons of interference", i.e., epistemic values that are the same for all members of the scientific community (Levi, 1960, p. 356).

Though there were still some debates about the proper role of values in science at the middle of the last century, the debate, at least in the USA, soon died out at the beginning of the 1960s in favour of the value-free ideal. While there were still some discussions ongoing about this topic elsewhere in the world,[5] it is worthwhile to consider why the value-free ideal spread so quickly in the USA

[3] Despite arguing for disinterestedness as an "ethos of science", Merton also sees science as embedded in a wider societal context in such a way that scientists have to consider the social and ethical implications of their work (Merton, 1973, pp. 267–278). Douglas also points out that science "'being value-free' is nowhere among the norms" (2009, p. 46) put forward by Merton.
Nagel (1961) also discusses the impact of values on sciences in an inner- and outer-scientific context in a variety of ways, though he does not regard the influence of values in science as far reaching as Rudner and Churchman do. A similar point of view is voiced by Hempel (1965); see also Douglas (2009, pp. 58–59).

[4] Other examples of prominent philosophers of the time who, Douglas notes, did not advocate for thinking about science as a fully value-free realm are, e.g., John Dewey and Rudolf Carnap (Douglas, 2009, p. 47).

[5] One might consider for instance, the second Werturteilstreit in Germany, which went on until the 1970s, and focussed on the role of values in sociology.

in the 1960s. In the post–World War II period the centre of discourse in philosophy of science, as much as science itself, shifted to the United States, which had, therefore, a global sphere of influence that did not end at the boarder. Further, it shows impressively how external (value-laden) pressure can influence the direction science and philosophy of science takes.

Douglas identifies two factors why the value-free ideal gained so much in popularity so quickly in the United States in the 60s. First, during the 1950s, the political situation in the USA had shifted in a way that put increasing pressure on all academics to distance themselves from anything that could be construed as support for communist ideas. Marxist philosophy traditionally sees science as situated in society so that the social and the scientific are interconnected. Fuelled by the McCarthy-era paranoia, many philosophers of science gave any position that could be misunderstood as political a wide berth.

This, Douglas notes, went hand in hand with another shift in philosophy of science already discussed in more details in Chapter 2.2: the separation of *context of discovery* from *context of justification*. Following Reichenbach's reasoning (1938), the former was deemed to be philosophically uninteresting and to be a topic of discussion for sociologists and psychologists but not philosophers.[6] Philosophy of science instead is supposed to focus on the logical justification of the result of scientific research. This distinction demands a restriction of the scope of research for philosophers of science, which was commonly interpreted to exclude a discussion about (social) values in science. The political situation made it attractive for philosophers of science in the USA to abandon any wide reaching, non-specific discussion about science in wider social context, Douglas argues, and instead "to professionalize their field, narrowing their expertise and focusing on a well-defined topic" (Douglas, 2009, p. 49).

The second factor Douglas cites as reason for the advancement of the value-free-ideal is the influence the publication of Thomas S. Kuhn's *The structure of scientific revolution* in 1962 had on philosophy of science. While the book had been hugely influential on philosophy of science in general, it also influenced the

[6] In this context it might be interesting to note that Don Howard (2006) argues that Reichenbach's DJ distinction has to be interpreted as a way for Reichenbach to directly distance himself from Otto Neurath and the idea that values have a legitimate role in science. According to Howard, Reichenbach's position is an attempt to exclude the question of values by reducing philosophy of science to pure logic, whereas Neurath saw certain values as an unavoidable byproduct of the underdetermination of theories in science.

debate about values in science, according to Douglas. In the book Kuhn places science in its own historical context but also situates science outside of society. The separation from society is what makes (natural)[7] science, for Kuhn, such a successful endeavour:

> the insulation of the scientific community from society permits the individual scientists to concentrate his attention upon problems that he has good reason to believe he will be able to solve. Unlike the engineer, and many doctors, and most theologians, the scientists need not choose problems because they urgently need solutions and without regards for the tools available to solve them. In this respect, also the contrast between natural scientists and many social scientists proves instructive. The latter often tend, as the former almost never do, to defend their choices of a research problem – e.g., the effects of racial discrimination or the causes of the business cycle – chiefly in terms of social importance of achieving a solution. Which group would one then expect to solve problems at a more rapid rate? (Kuhn, 1962, p. 163)

By the mid-1960s the value-free ideal had truly become mainstream. Douglas writes that even in post-McCarthyism times, the ideal of value-free science was very attractive to science and philosophy of science alike for several reasons. For instance, the notion of the intrusion of values into science might 'bring back bad memories' of periods in the history of science where unwarranted outside forces interfered (see also Rudner, 1953, p. 6), or there might be concern that "science will lose its general public authority if a role for social or ethical values is admitted" (Douglas, 2009, p. 79).

The discussion about the role of values in science only gained momentum with the emergence of feminist philosophy of science. Feminist philosophers (for

[7] Kuhn distinguishes here between natural and social sciences and sees the latter much more situated in a social context than the former. This, according to Kuhn, already transpires in the way that the training for future scientists is structured in the different disciplines: while the social scientists are required to study the original sources of previous scholarly disputes, where they learn to see different perspectives and arguments, the student of natural sciences is presented with condensed versions of research results from textbooks. Only advanced students actually study research papers directly. This "rigid education", according to Kuhn, prepares the young scientists optimally (in a period of normal science) for a (professional) life of puzzle solving (Kuhn, 1962, pp. 164–165).

example, Helene Longino, Donna Haraway and Sandra Harding) raised concern that there is no such thing as an objective, value-free "gaze from nowhere" (Haraway, 1989, p. 581).[8] Rather, they argue, upholding the value-free ideal would only hide the actual value judgements that come into play when doing science. Instead, feminist philosophers have promoted an honest handling and open communication of (potential sources of) value-influence in science. A popular proposal made by feminist philosophers of science to counterbalance inadvertent value-laden background assumptions is a more pluralistically and diverse organised science community (Longino, 1990). Although the feminist criticism of the value-free ideal was first met with reservation by conventional philosophy of science (see for example Kitcher, 1993), in the last two decades, the thinking that value judgements are inevitable in science has come back into the mainstream of philosophy of science. Particular the rediscovery of Rudner's 1953 paper has reactivated research interests into the question if and when values are an appropriate feature of science (e.g., Carrier, 2013; Douglas, 2009; Wilholt, 2009).

3.1.1.2 Epistemic versus non-epistemic values

With the rise of the value-free ideal another distinction also arose: that between *epistemic* and *non-epistemic* values. While many philosophers argued for value-free science, they also recognised that scientific theories on their own are underdetermined (see Chapter 3.2.2). To choose between two competing theories, more than empirical evidence is required. The same empirical evidence can support several even contradicting theories. Scientists, thus, need something more to make decisions between two or more equally well-established theories; some kind of value judgement is needed. Some philosophers have, therefore, suggested to make a distinction between appropriate, inner-scientific values and those values that come from an outer-scientific realm that the advocates of the value-free ideal try to keep out of science. Going back to Ernan McMullin (1982), these two types of values are often referred to as *epistemic* and *non-epistemic* values. More specifically, McMullin refers to five criteria defined by Kuhn, which are to fulfil the role of epistemic values in science: accu-

[8] Ironically, many of those who first advocated for a value-free science at the turn of the last century made the (arguably value-laden) judgement that objectivity and value-neutrality were male attributes and women were, therefore, deemed not suitable for science (Proctor, 1991, p. 119).

racy, consistency, scope, simplicity and fruitfulness (Kuhn, 1977, p. 322).[9] Proponents of this distinction argue that epistemic values are distinct from non-epistemic ones because they "are presumed to promote the truth-like character of science" (McMullin, 1982, p. 702). Though they are "normative principles [...] given an initial commitment to these principles, the scientist need not and should not let his values, attitudes, and temperament influence his inferences any further" (Levi, 1960, p. 346).

Thus, the value-free ideal as it manifested itself in the early part of the second half of the last century is not a complete rejection of any kind of values, as Douglas notes:

> The ideal that has held sway since 1960 is a complex one. It does not hold that science is a completely value-free enterprise, acknowledging that social and ethical values help to direct the particular projects scientists undertake, and that scientists as humans cannot completely eliminate other value judgements. However, the value judgments internal to science, involving the evaluation and acceptance of scientific results at the heart of the research process, are to be as free as humanly possible of all social and ethical values. Those scientific judgements are to be driven by values wholly internal to the scientific community. Thus the value-free ideal is more accurately the "Internal scientific values only when performing scientific reasoning" ideal. (Douglas, 2009, p. 45)

By introducing this division between 'good' epistemic values and 'bad' non-epistemic, social values, the image of science as a space free from at least external scientific influences appears to stay intact, even if one recognises the issue of underdetermination.

But in what sense are these epistemic values actually distinct from non-epistemic social, ethical or political values? Critics of this dichotomy have argued that many of the values that are traditionally seen as purely epistemic do not strictly say anything about the truth of a theory (Laudan, 2004). For instance, a theory can be true even when there is an alternative one with a wider

[9] Although Kuhn states that these five criteria are not exhaustive, they are often treated as the canon of epistemic values by philosophers of science. Rooney shows how philosophers make slight differences in how they define and name those criteria but finds that the "fact that there is no clear consensus about what is included among the epistemic or constitutive values does not overly concern many of those who make the distinction" (1992, p. 14).

scope. Similar things can be said about the values of *simplicity* or *fruitfulness*. That does not mean that these kinds of values are of no worth to science. There is more to a 'good' theory than accuracy, as Laudan with reference to Bas van Fraassen points out:

> Bas van Fraassen famously argued that a theory does not have to be true to be good. We can add to that dictum a new twist: a theory does not have to be false to be bad. A theory may be bad because it fails the test of possessing the relevant nonepistemic virtues. In other words, we expect our theories to do much work for us, work of a sort that most merely true statements fail to do. (Laudan, 2004, p. 19)

Nor are epistemic values as universal as their advocates claim. Helen Longino (1990, pp. 83–102, 2002, 2008) argues that one could easily imagine a set of alternative epistemic values, which would also fill the gap left by the underdetermination of theories. One might, for instance, substitute the values of *simplicity* and *scope* by *heterogeneity* and *mutuality*. There is, after all, "no prior reason to think the universe simple, that is, composed of very few things" (Longino, 2008, p. 73).

Besides *empirical adequacy*, Longino proposes feminist philosophers might advocate for alternative epistemic values such as *novelty*, *heterogeneity*, *mutuality* and *decentralisation of power* because they support feminist objectives in science. Longino provides several examples from the history of science to argue which specific values are observed can have actual social consequences. Such instances are medical research only done on white males or economic theories of the household assuming patriarchal structures, both adhering to an ideal of simplicity (Longino, 2008, pp. 74–75).

Therefore, argues Longino, these kinds of values – the traditional epistemic values and any kind of alternative values – are actually heuristics. These heuristics are specific to particular scientific communities and might shift over time when they no longer serve their purpose. Thus, the alternative values Longino proposes are not feminist in themselves, even though feminist might favour them but "subordinat[e] to a broader cognitive goal" (2008, p. 78). Ergo, there is nothing special about so-called 'epistemic' values:

> The feminist and traditional virtues are on par, epistemologically. Both have heuristic but not probative power. As heuristics, they help an investigator identify pattern or order in the empirical world. They are often transmit-

> ted as part of an investigator's training, as part of the common, taken-for-granted-background. (Longino, 2008, p. 74)

Hence, opponents to the value-free ideal have argued that the alleged epistemic values are not purely epistemic in nature but emulate current dominant social values. Feminist philosophers, for instance, have demonstrated in case studies from the fields of biology, primate studies, economics and medicine (amongst others) how androcentric values are sometimes 'hidden' behind the apparent 'objective' epistemic values.[10] Because of this some philosopher such as Laudan (2004) and Douglas (2009) have argued that the term *epistemic* should be dropped completely to describe values that help scientists in their reasoning process and that the more appropriate word for these kinds of values would be *cognitive*. *Cognitive* then refers to "those aspects of scientific work that help one think through the evidential and inferential aspects of one's theories and data. Taking the label *cognitive* seriously, cognitive values embody the goal of assisting scientists with their cognition on science" (Douglas, 2009, p. 93). Thereby, cognitive values function as an "insurance policy" (Douglas, 2009, p. 107); they increase the likelihood to find possible mistakes in the reasoning process.

But what about those virtues that are actually truth-conductive? Laudan argues that 'true' epistemic values are a small subsection of the cognitive values (2004, p. 19). Douglas, however, suggests actual epistemic virtues such as *empirical adequacy* or *internal consistency* should be viewed not as values but as necessary criteria for any kind of scientific theory (2009, p. 94). For example, it seems difficult to imagine a satisfying scientific theory that is not internally consistent.

This way of defining the role of values in science paints a different picture than the value-free-ideal. There is no longer a clear separation between 'acceptable' epistemic values and 'not acceptable' non-epistemic values. Douglas argues one should instead imagine the different types of values as different areas on a landscape that might intersect in certain places (Douglas, 2009, p. 91).

In my opinion Longino, Douglas, Laudan and others have raised valid concerns in respect to the dichotomy of *epistemic* and *non-epistemic* values. But one might as well question whether the term *cognitive* is not equally badly defined and far-reaching. Moreover, one might even ask whether it makes much sense to make

[10] See, e.g., Haraway (1989); Keller (1985); Longino (1990). For an overview, see also Douglas (2009); Longino (2008); Rooney (1992).

this distinction at all if one takes Longino's claim serious that cognitive values might actually be based on concepts external to science. I will, however, not discuss this any further here as that controversy would lead us astray.

In order to avoid confusion over what the terms *epistemic* and, therefore, *non-epistemic* values actually refer to, I will in the following use the term *social and ethical values*. The allegations that climate scientists are biased and the subsequent philosophical debate about the role of values in climate modelling focuses exactly on those values; the kinds of values that proponents of the value-free ideal fear would 'contaminate' scientific research. The question what actually the appropriate place for values in climate science is and how severe the value-ladenness of climate science actually is centres on *social* and *ethical* values. The role of *epistemic* or *cognitive* values in climate science, on the other hand, is normally not disputed. Though there are differences between *social* and *ethical* values (Douglas, 2009, pp. 92–93) in general, I will not discuss these any further here. As this chapter will show, the practice of climate modelling involves a myriad of epistemically not fully constrained decisions and it is often impossible to determine retroactively if and what values were relevant in making those decisions. Thus, the question will be to what extent it can be argued that these kinds of possible value-laden judgements are not inappropriate (as the value-free ideal asserts) rather than what constitutes the specific value. For this reason I will refer to *social values* in short for *social and ethical values* as the kind of values that are contrary to cognitive values and according to the value-free ideal do not have a place in scientific reasoning processes.

Much of the current discussion about the relevance and unavoidability of social values in science, including climate science, is built on what is known today as the *argument from inductive risks*, specifically two forms of the argument made by Rudner (1953) and Douglas (2009). Which is why I will now discuss the inductive risk argument in more detail before turning to the specific issues raised by philosophers of climate science. In the following discussion of Rudner's and Douglas' remarks on value judgements in the context of inductive risks, I will specifically use the term *ethical* value where it supports the argument – contrary to the later discussion of the situation in climate modelling, where such a differentiation is, as said above, not particularly helpful, (otherwise I will continue to use the short version of *social values*, where both types of values are concerned).

3.1.2 Inductive risks and social values

Carl Hempel, who first introduced the term, defines inductive risks as the risk of accepting an empirical law to the extent that "the presumptive law may not hold in full generality, and that future evidence may lead scientists to modify or abandon it" (Hempel, 1965, p. 92). In these situations, opponents of the value-free ideal contend that scientists sometimes cannot but must make social-value judgements.

In his widely discussed 1953 paper "the Scientists qua Scientists makes value judgments" Richard Rudner argues that contrary to what the value-free ideal proposes – that there is an appropriate, even necessary role for social-value judgements within science, under the assumption that accepting or rejecting hypotheses is a quintessential part of the scientist's work.[11] In a nutshell the argument goes like this: when one accepts this premise and also assumes that ultimately a scientific hypothesis can never be fully verified, then one also has to assume that "in accepting a hypothesis the scientist must make the decision that the evidence is *sufficiently* strong or that the probability is *sufficiently* high to warrant the acceptance of the hypothesis" (Rudner, 1953, p. 2). In those cases where the research objective has a wider social application, judging whether or not a hypothesis is sufficiently proven has to be done under consideration of social values. Rudner demonstrates this with the now often quoted example of the testing of drugs versus belt buckles before they are released for sale. We require a much higher standard before we accept the hypothesis of a drug containing a toxic ingredient, only in such a quantity that it is still safe to use than we require of a load of belt buckles based on a sample size because the stakes in the first case are so much higher. Put slightly differently, compared to the consequences of falsely accepting the hypothesis that a drug is safe to use, a not properly working belt buckle might be annoying or embarrassing but has no (or only under exceptional circumstances) possible deadly consequences. Thus, a much more thorough testing is required when it comes to medicine than clothing accessories. This difference in requirements for testing standards arises not because of epistemical but ethical considerations. For this reason, Rudner concludes that the scientist cannot avoid value judgement in their role as a scientist.

[11] Similar opinions have been voiced by others at the time, such as Churchman (1948) and Frank (1953). However, Rudner's paper has been the one discussed the most in subsequent years.

One might now be inclined to argue that the scientist can nevertheless hand over the responsibility of making value judgements to the public. This position, commonly ascribed to Richard Jeffrey, who claims that the scientist themself does not have to accept or reject a hypothesis but merely has to disclose the probability for one. According to Jeffrey, the scientist, after all, could not possibly have insight into all the possible consequences of their acceptance or rejection of a hypothesis:

> One cannot, by accepting or rejecting the hypothesis about the polio vaccine, do justice both to the problem of the physician who is trying to decide whether to inoculate a child, and the veterinarian who has a similar problem about a monkey. To accept or reject that hypotheses once for all is to introduce an unnecessary conflict between the interests of the physician and the veterinarian. (Jeffrey, 1956, p. 245)

Thus, following Jeffrey, the scientists are in no way forced to make judgements about the acceptability of a hypothesis. This is something that can be passed on to those who apply the information they get from the scientists in practice.

Rudner, on the other hand, argues that even just stating a probability for a hypothesis requires value judgements, as stating a probability p for the occurrence of a hypothesis H requires *“the acceptance by the scientist of the hypothesis that the degree of confidence is p“* (1953, p. 4). That is, coming to the conclusion that the probability of H is p requires accepting a further hypothesis H'. Therefore, a common counterargument to Jeffrey's position is that the problem of the value judgements is just transferred to another level.

The disagreement between Rudner's and Jeffrey's position regarding the need for scientists to make value judgements is often framed as a disagreement over the question of whether or not probabilities attributed to hypotheses are actually something that can be accepted or if they constitute a degree of belief in a conventional Bayesian sense (Steel, 2015; Winsberg, 2012). According to the latter position, personal probabilities are not something to be accepted but something one has, which may shift in light of new evidence but does not usually involve a conscious decision process. From this point of view the counterargument to Jeffrey that noting probabilities requires acceptance of those does not hold up. However, it seems questionable to what extent scientists can actually be described as perfect Bayesian actors. Steel (2015), for instance, points out that scientists often hold vague degrees of belief. In Chapter 3.1.3 we will return to this issue with a climate-science specific argument first made by Winsberg (2018, 2012).

As discussed in Chapter 3.1.1.1, the position of (social-)value judgement having no place in the scientist's everyday life as scientists, prevailed in philosophy of science in the subsequent years. In the last few decades, however, Rudner's paper emphasising the importance and unavoidability of social-value judgements in cases of inductive risk assessment in science has seen a resurging interest. Specifically noteworthy here is Heather Douglas's (2000) account of how scientists often are confronted with a number of methodological decisions that require social value judgements even before they accept or reject a final hypothesis. While Douglas does not, as has been pointed out, directly refute the Bayesian interpretation of Jeffrey's argument (Steel, 2015; Winsberg, 2012, 2018, pp. 135–136; Parker and Winsberg, 2018), it is still worthwhile to take a look a Douglas' reasoning for two reasons. First, if one takes scientists not to be perfect Bayesian actors, then it indicates that value judgements are a common and necessary element to decision-making processes in science way before scientists have to decide whether or not to accept the hypothesis. Second, Douglas also concludes from this that one can assess the appropriate place of value judgements in science by distinguishing between a direct and indirect role of values in science, a question we will also return to.

3.1.2.1 Social values and methodological considerations

Scientists do not only deal with inductive risks and thereby value judgements when evaluating hypotheses at the end of a scientific project. Douglas identifies three stages at which, she argues, social values can and do have a legitimate role in science-internal processes due to considerations of inductive risks[12] – which is not limited to the final evaluation of the hypothesis:

> If one follows the general schema of the methodology from a scientific research paper, significant inductive risk is present at each of the three "internal" stages of science: choice of methodology, gathering and characterization of the data, and interpretation of the data. At each point, one can make a wrong (i.e., epistemically incorrect) choice, with consequences following

[12] Note that Douglas' definition of the term *inductive risks* here is rather board and goes beyond the risks of accepting or rejecting a hypotheses, but refers to errors more generally, i.e., making a "epistemically incorrect" decision at "'internal' stages of science" (Douglas 2000, p. 565). Harvard and Winsberg (2022) criticise this broad definition of inductive risks and argue that differentiating between *inductive risks* and *representational risks* is useful, particular in the context of scientific modelling (see Chapter 3.1.3.1).

> from that choice. A chosen methodology assumed to be reliable may not be. A piece of data accepted as sound may be the product of error. An interpretation may rely on a selected background assumption that is erroneous. Thus, just as there is inductive risk for accepting theories, there is inductive risk for accepting methodologies, data, and interpretations. (Douglas, 2000, pp. 564–565)

Douglas rejects the assumption that in these cases the task of making value judgements can simply be handed over to the public. She illustrates this with the example of a case study of animal testing and the risk of dioxin inducing cancer in rats.

First of all, Douglas argues that in the specific case study one methodological decisions scientists have to make concerns the level for statistical significance (2000, pp. 565–569). As the control group of rats will also show a natural amount of cancer, the scientist must decide on a standard for statistical significance, beyond which the amount of cancer found in a rat population is considered the result of the exposure to dioxin. If they choose not to go along with conventions of that particular research field,[13] Douglas notes, scientist then have to consider the consequences of false positive (that is falsely accepting a hypothesis) or false negative (erroneously rejecting a hypothesis) errors when defining the level for statistical significance. Depending on how low or high the standard for statistical significance is set, one risks more false positive or false negative results. When the toxicity is wrongly overestimated, it may have a negative impact on the affected industries, whereas when scientists underestimate the possible toxicity, serious consequences for public health may arise. Therefore, the decision to go one way or the other is a question of trade-offs. Douglas emphasises that one can only reduce both the risk of false negative and false positive errors under a significant increase of costs, that is, in this case improving the experiment by increasing the number of research objects. Thus, it comes down to the question how the scientists value the possible consequences of both options:

> In finding the appropriate balance between false positive and false negative errors, we must decide what the appropriate balance is in the consequences

[13] It seems nevertheless reasonable to note that even when scientists follow these conventions, one might as well argue that in those cases scientists do make a value judgement to go along with such conventions.

> of those errors: overregulation and underregulation. Selecting an appropriate balance will depend on how we value the effects of those two consequences […]. Finding the balance requires, among other things, weighing the non-epistemic valuations of the potential consequences. (Douglas, 2000, p. 568)

That is, in situations of methodological uncertainty where the decision will have political and social implications scientists cannot but refer to social values, Douglas argues.

Secondly, when it comes to the characterisation of data, evidence is rarely unambiguous. In these instances and when the results have consequences for public safety, scientists likewise have to consider the risks of false positives and false negatives, Douglas notes (2000, pp. 569–572). In the dioxin cancer studies Douglas examines, there was a significant number of cases where the scientists did not agree if the rat liver slides showed cancerous lesions. Different groups of pathologists who had evaluated the same samples at different points in time came to different conclusions how those borderline cases should be classified. In some situations, as Douglas points out, these types of discrepancies can be circumvented to a certain degree by letting the pathologist examine the tissue sample 'blind', so that they do not know whether the samples come from the rats exposed to dioxin or the control group. Therefore, when scientists have an evenly distributed tendency to false positives (or negatives), the errors in judgement should (in theory) balance each other out. But this approach would not work for most borderline tissue samples in Douglas' case study as the liver tissue samples of rats having been exposed to a high level of dioxin would also show signs of acute liver toxicity, which the experienced pathologists evaluating the samples would recognize. Thus, Douglas concludes, inductive risks and associated social-value considerations also do play a role in data assessment in this particular case and beyond:

> This case demonstrates that there is inductive risk in how one applies categories used in data characterization and that such inductive risk can be linked to non-epistemic consequences. […] The consequences of the errors are identifiable and need to be weighed in order to determine which errors are more acceptable. In other cases, inductive risk may be present in the selection of the categories to be used as well as the application of the categories in the characterization of the data. In addition, judgments are made in science concerning whether to keep data or whether to discard the data as unreliable. At all these decision points, there is the risk of error, and

> with that risk, the need to consider both the epistemic and non-epistemic consequences of error. (Douglas, 2000, p. 572)

The third instance, where scientists have to consider inductive risks in their daily work, is the interpretation of data, according to Douglas (2000, pp. 573–577). In the case of the dioxin-study, there was considerable disagreement whether the result should be interpreted as there being a specific threshold for the dose of dioxin after which it causes cancer or whether the response increases consistently with the dosage. Thus, Douglas argues scientists have to take the inductive risks into account and weigh the consequences of potential errors in their judgement. Depending on whether one chooses a *threshold model* or a *linear extrapolation model*, the acceptable dosage for human consumption will be set differently. The consequences of a linear extrapolation model are usually stricter regulations than of a threshold model. Which one is chosen has consequences for the general public and the industry more specifically, as a wrongly chosen threshold model will most likely have negative consequences for public health whereas if it turns out that one erroneously adopts a linear extrapolation model industries will most likely be overly regulated.

Thus, in the case of socially sensitive research, scientists have to make social-value judgements at different stages in the science internal process, concludes Douglas. Further, in those cases, where inductive risks are at play "value-free science is inadequate science" (Douglas, 2000, p. 559). And contrary to Jeffrey's claim, scientists are the most competent and often only option for making these decisions, Douglas claims:

> The most important reason is that it is doubtful anyone could fully take over this function for scientists. Because science's primary goal is to develop knowledge, scientists invariably find themselves in uncharted territory. While the science is being done, presumably only the scientist can fully appreciate the potential implications of the work, and, equally important, the potential errors and uncertainties in the work. And it is precisely these potential sources of error, and the consequences that could result from them, that someone must think about. The scientists are usually the most qualified to do so. (Douglas, 2009, pp. 73–74)

In an attempt to clarify the appropriate role for social values in science, Douglas also differentiates between the direct and indirect role of values (2009, pp. 95–108). At the early stage of a research project, value judgements often have an (acceptable) direct role, Douglas points out. There are, e.g., legitimate reasons for ruling out certain methodological approaches for ethical reasons from the beginning. We do not, for instance, endorse certain kinds of experimentations on human beings for ethical reasons. In a similar fashion, ethical or social values may be relevant when determining the objective of a research project. For example, governments might be more inclined to fund those projects which have distinct social relevance. Douglas argues, in these sort of situations, at the beginning of a scientific project, social values may take a direct role in such a way that they "determine our decisions in and of themselves, acting as standalone reasons to motivate our choices" (2009, p. 96). Values having a place in establishing research objectives and setting ethical boundaries when it comes to methodology, are widely accepted – even by proponents of the value-free ideal – because these decisions are seen as still taking place at a stage of the research project that has a 'pre-scientific' character.

Much more contested is what Douglas identifies as the indirect role of values. It refers to the role of values in science we have primarily discussed so far. Douglas applies this term to those instances during an ongoing research project when scientists have to make decisions under uncertainty. They help review whether there is sufficient evidence considering the specific circumstances, make decisions and weigh the consequences of potential errors in judgement in the way discussed above. In this form, the role of values is contingent upon the specific evidence at hand, Douglas contends.[14] When new evidence reduces the uncertainty, it also reduces the need and the place for values.

Thus, Douglas argues, depending on what stage a scientific research project is at, values may take up different roles. At an early stage, values can legitimately direct our choices by putting value on it in itself, whereas during the research project values should only take an indirect role and aid scientists when they are facing uncertainties.

However, these roles are not as clear-cut as they might seem at first glance, Douglas concedes. Under certain circumstances, direct value judgements might be inappropriate, even at an early stage of a scientific project. These, notes Douglas, are those cases where "a direct role […] undermines the value of science itself" (2009, p. 101). This might be the case when objective and

[14] That the role of values should be limited to an indirect role, once a research project is under way, also holds for cognitive values, according to Douglas (2009, pp. 107–108). This will not be discussed in great detail here because social, not cognitive values are at the centre of the argument concerning value judgements in climate science.

methodology predetermine the result of a research project. Furthermore, Douglas concludes that under certain (exceptional) circumstances, it may be appropriate for values to interfere directly in the science-internal process. Such might be the case, when scientists have to adjust their methodology, because it turns out that the methodology chosen at the early stage of the research project is in fact ethically not acceptable. We will return to the question to what extent it makes sense to distinguish between a direct and indirect role of values in science at the end of Chapter 3.1.

3.1.3 Social values in climate science

Once we turn our attention to scientific disciplines which have to deal with additional epistemic challenges coming from the high complexity of the systems under investigation, such as climate science, it becomes clear that the role of values within the scientific process reaches even further than inductive-risk assessments. In such cases, contrary to what the proponents of the value-free ideal envisioned, the significance of social values cannot be reduced to the role of setting goals at the beginning of the model-building process. Nor can the role of social-value judgements at internal stages of the scientific process be cut down to, as Douglas argues, decisions under uncertainties. Rather, as Winsberg has shown, the possibility of social value-judgements lies deep within the "nooks and crannies" (2012, p. 132) of climate science.

What differentiates the cases of social values in science discussed here so far from climate science are the specific epistemic challenges rooted in the high complexity of both system and models which entails a great deal of epistemically not fully constrained decision making. Because of this, as we will see, when it comes to the relevance of value judgements, "predictive preferences" (Winsberg, 2012) gain in significance and "representational risks" arise (Harvard and Winsberg, 2022). In the following I will discuss the consequences of this for the ways that social values might interfere in the model-building process. It will be shown that the complexity of the climate models does not only make it impossible to rule out that some science-internal decisions were made under considerations of social values but even to retroactively (fully) disclose them. In fact I will claim, this complexity is the reason why we do not have to fear that bias and wishful thinking could impact the models in an epistemically untoward way.

3.1.3.1 Unconstrained decision making, predictive preferences and cost restrictions

As argued before (see Chapter 2.1), there is no such thing as a fixed construction manual for a climate model. On the contrary, there is a plurality of models concerning scale, complexity and objective (for an overview see Parker, 2018). As has been pointed out, considering all idealisations and trade-offs, the quality of these kinds of models can only be assessed with respect to their purpose (Chen et al., 2021; Parker, 2009). But even with models possessing a similar set of goals, scale and complexity, there are numerous ways to construct a global climate model. On a macroscopic level, Alexander and Easterbrook (2015) have shown that there are different modelling traditions in Europe and America, which are both epistemically equally well justified (see Chapter 2.1). The emergence of these kinds of epistemically unforced methodological questions are not singular events in the process of 'assembling' these kinds of computer simulations. It is not an uncommon occurrence in process of model construction that there are several different options how to represent one and the same climate mechanism within the model depending on the particular objectives and underlying modelling 'philosophies'. In this context scientists are also often faced with the question to what extent and in what way to include specific processes. How these questions are answered depends on the purpose of the model but also on cost-benefit deliberations.[15] For instance, one might imagine a hypothetical situation where three different modelling groups have to decide how to implement a climate-relevant process into their model, which can be represented either resolved or parametrised. One modelling group, for example, might choose to represent a specific process in their new model by relying on a parametrisation that they are well acquainted with, have used and tested in a previous model because they are content with the performance of the model with said parametrisation. A different research group might decide that it is worthwhile to invest in increasing the resolution of their new model so that it is possible to integrate that specific process in their model directly in a resolved way. The scientists conclude that this will hopefully result in a physically more accurate representation of the process. A third modelling group, however, might make the decision to improve the existing parametrisation in

[15] 'Cost' has to be understood in the broadest sense here, see below.

their current model because they think that that process is more effectively represented in the model in form of a new parametrisation.[16]

Similarly, it is widely acknowledged among climate scientists that there is a variety of ways of how to tune a model or model components. Tuning is the fixing of certain, otherwise not very well constrained parameter so that the overall model result fits better with scientists' expectation based on observations and expertise. Although there are some conventions, as it is not possible to tune a model perfectly in respect to every variable, differences in priorities and in well-established modelling cultures at individual institutes also influence what approach to the tuning process is taken (Chen et al., 2021, pp. 217–218; Mauritsen et al., 2012). Even though scientists will have good reasons for choosing the tuning method and objectives that they do, they are also aware that their decision are not strictly epistemically constrained (Hourdin et al., 2017).[17]

In practice, how these decisions are made is often also dependent upon the modelling culture at different research institutes and the specific histories of the models. However, it should be emphasised here that these are decisions where the scientists have good (epistemic and methodological) reasons for deciding the way they do. But if they had chosen one of the other paths, because of

[16] In practice, as Helen Guillemot (2017) has shown, modelling groups only rarely invest in improving existing parametrisations because the costs in terms of time and effort are estimated to be too high compared to the benefit of the outcome. Even though a new parametrisation in isolation might seem as an improvement, due to compensating effects and tuning, it will inevitably perform worse than the old one when first integrated into the model, which means additional work. See also Chapter 3.3.3.2 and Chapter 2.1.

[17] Historically, tuning has not been an issue that has been discussed much within the climate science community. Hourdin et al. (2017) point out two possible reasons for this. On the one hand, tuning may be considered to be somewhat 'unscientific' and "more engineering than science, an act of tinkering that does not merit recording in the scientific literature" (2017, p. 590). On the other hand, there may be concerns that emphasising the necessity of tuning may give 'ammunition' to climate change sceptics. So that the climate science community may see the whole process of tuning "indeed as an unspeakable way to compensate for model errors" (Hourdin et al., 2017, p. 590). However, in the last decade there have been several attempts to bring the actual reasoning process behind different tuning strategies to the forefront (Schmidt et al., 2017). Most notably, in the widely discussed paper from Mauritsen et al. (2012) the authors explore as a case study what the effects are of different choices made in the tuning process of their model.

different objectives or modelling cultures, that choice would have been equally well justified.[18]

Considering the specific epistemic challenges of climate-model development, particularly the difficulties in accessing analytical understanding (see Chapter 2.1) and the long timeframe, these are not necessarily decisions that can be made fully at an early stage when a new model is initiated. As will be argued in the following, these are choices that have to made and (re-)assessed continually during the process of constructing and evaluating climate models.

In the context of climate science the role of value judgements are mostly discussed in the context of what Eric Winsberg has termed "predictive preferences" (2012, p. 131, 2018, p. 138). As there is no such thing as a perfect climate model representing every aspect of the global climate equally well, climate models echo predictive preferences climate scientists have for specific variables or processes that they consider to be more significant for their research questions (Tebaldi and Knutti, 2007, pp. 2045–2055). Trade-offs have to be made with respect to which aspects of the climate system are to be prioritised, not just in respect to where one decides to invest time and money but also purely on grounds of the intricate characteristics of the model building process. Considering that the model show path dependency, specific modelling choices will restrict what further options are available and impact how well other aspects of the climate can be represented in the model. The setting of priorities is also a necessary feature of climate model tuning (Mauritsen et al., 2012).[19] One cannot tune the perfect model; a model can only be tuned well with respect to certain aspects. There are again trade-offs to be made to the

[18] This is not unique to climate science; it also affects other scientific disciplines using computer simulations of a similar complexity. A similar situation is, for instance, described by Ruphy (2016, pp. 100–101) in relation to the use of computer simulations in astrophysics.

[19] Hourdin et al. describe what kind of different objectives these might be and what influences them: "different models may be optimized to perform better on a particular metric, related to specific goals, expertise, or cultural identity of a given modelling center. Groups more focused on the European climate may give more importance to the ocean heat transport in the North Atlantic, whereas others may be more concerned with tropical climate and convection. Some groups may put more weight on metrics that measure the skill to reproduce the present-day mean climatology or observed modes of variability, while others may privilege process-oriented metrics targeting processes that are believed to dominate the climate change response to anthropogenic forcing" (2017, p. 592). Note that possible tuning goals listed here do not just concern

extent that tuning models to improve the representation of specific features of the climate often goes hand in hand with a decline in the model performance in respect to other aspects of the climate (Hourdin et al., 2017, p. 596). These value judgements concerning what to prioritise when tuning a model also cannot be circumvented by relying on algorithmic (so-called *objective*)[20] methods to find optimal parameters for a specific target, because it still requires the scientists to set goals for the tuning process. As Hourdin et al. point out: "An objective algorithm merely identifies those parts of the procedure that require the subjective scientific expertise of the modeler" (2017, p. 594).

There are further ways in which social-value deliberations might influence methodological decision-making in climate science. Often these kinds of decisions are determined by *considerations of costs*. Costs have to be understood here in a broad sense, it does not only include financial deliberation but also, for instance, questions in respect to time and effort put into developing a model or specific part of the model.[21] These considerations are similar to predictive preferences but are different in the primary goal, though a finite number of resources will eventually also lead to predictive preferences.

Harvard and Winsberg (2022) argue that in the context of computer modelling one ought to distinguish between inductive risks and representational risks. Representational risks go beyond the risk of upholding a false fact. Harvard and Winsberg emphasise that representational decisions are not questions of right or wrong but whether or not an adequate choice for the intended purpose and considering all relevant epistemic agents is made. This is an important distinction that does not just pertain to the here discussed occurrence of unforced decision making in climate modelling. Harvard and Winsberg note that in the context of complex computer simulations in general it is a well-established insight that they often include elements that do not have a direct representation in the 'real' world. It is not even uncommon in complex computer simulations that a 'false' parameter might be the right choice as it adequately compensates for inaccuracies elsewhere in the model. One example Harvard and Winsberg point out from climate modelling are parameter values for cloud formation,

pure scientific research questions but may also serve distinctly socio-political purposes concerning dealing with anthropogenic climate change.

[20] See Chapter 2.3 and Chapter 3.4.3.

[21] In science time and effort of can, of course, again be translated into financial costs (see also Knorr-Cetina, 1981, pp. 40–41).

which are chosen to balance energy leaks in climate models at the top of the atmosphere (2022, p. 15).

One might also ask the question to what extent different modelling cultures influence how epistemically unforced decisions are made. Different modelling groups or institutes can develop different traditions concerning central questions in the modelling process (Skelton et al., 2017). Besides the differences between European and American model structures (Alexander and Easterbrook, 2015) already discussed here climate scientists have also stressed that there are different approaches to tuning at different institutes (Hourdin et al., 2017; Mauritsen et al., 2012). There can also be diverging attitudes towards basic ideas about the future of climate modelling, e.g., whether to invest in better parametrisation or in reducing the grid size so as to reduce the dependency of models on parametrisations (Guillemot, 2017). This is similar to different 'lab cultures' which are traditionally attributed to groups of scientists working in laboratory settings (Knorr-Cetina, 1999; Latour and Woolgar, 1979).

At first glance these modelling cultures seem to resemble *cognitive* values, insofar as they seem not to be affected by non-science related assumptions. However, just as cognitive values might as well be fundamentally grounded in political or ethical ideals (Longino, 2008), it does not seem too far-fetched to question if this might not also be the case for these modelling cultures. The belief that scientists should invest more time and energy into improving parametrisation, for example, instead of being hell-bent on reducing the grid size of models could also be influenced by the hope to thereby provide better, policy-relevant results considering the time constraints.

While philosophers of science have discussed the role of values in climate modelling extensively, one question that has so far has seen much less attention is whether the collection and creation of observational data (sets) are similarly affected by social value deliberations. But it is to be assumed that the situation here is not much different to that in climate modelling. Observational data are often understood by laypersons as some kind of 'objective' benchmark against which the quality of a theory or a model can be assessed. However, as will be discussed in more detail in Chapter 3.2.3, climate data, just like models, are impacted by a wide variety of uncertainties and inaccuracies. Therefore, a great deal of processing in terms of filtering and homogenising has to be done to create global data sets. This comes hand in hand with some degree of methodologically not fully constrained decision making (Parker, 2018). As, for example, the case of satellite data shows (Chapter 3.2.3.1.2) there is some wiggle room in how to interpret this kind of data, the possibility of some inductive risk consideration taking place does not seem to be too far of. Further, Brönnimann and Wintzer (2018) point out the context-dependency of climate data. That is, "climate data products carry imprints of social, political and economic contexts" (Brönnigmann and Wintzer, 2018, p. 4) of the circumstances under which they were created. One place they note this can be observed is in the historical inequality in climate data coverage, which they argue "is not just a data problem, but also one that affects climate justice" (Brönnigmann and Wintzer, 2018, p. 4). As will be further explicated in Chapter 3.2, climate models are not fully theoretical constructs but are "data-laden" (Edwards, 1999), so what data is collected and available also has an impact on the model-building process. Thus, it has to be assumed that the decisions made in the process of the creation of data sets (which do, as can be derived from Brönnimann and Wintzer's analysis often resemble predictive preferences) are anchored deep within the "nooks and crannies" (Winsberg, 2012, p. 130) of climate modelling.

Now one might be inclined to further discuss in what particular ways social values are relevant in specific climate-science internal processes and what their specific impact is. But, in my opinion, that would be somewhat missing the point. What is of relevance here is not which specific values influence climate modelling, but that there are, as Winsberg has pointed out, literally "thousands of unforced methodological choices" (2012, p. 130), which require to set priorities that cannot be determined purely on the basis of epistemic considerations alone.

3.1.3.2 Non-traceability

In the debate about values in science, it has been argued that in those cases where scientist have to make judgements on the basis of social-value deliberations scientists should take care of communicating what went into these decisions as explicit as possible to policy makers and the general public to ensure the integrity of science (Douglas, 2009, p. 136). Against the backdrop of the vast number of epistemically unforced decisions and the different ways in which social values may play a role in the construction of a climate model, the question is whether this call for disclosure of all possible value-laden assumptions can still be met.[22] What complicates this even further is the fact that complex, global climate models are not built by just one scientist or even one research group with a fixed small number of scientists. Even though the traditional ideal of a scientist is that of a lonely man working in his lab or at his desk, the notion of science being a community effort is well established in philosophy of science. What is relatively new to science and has so far not been examined by philosophers as much is what has been called "radically collaborative research" (Huebner et al., 2017; Kukla, 2012; Winsberg et al., 2014), where the research is shared between many different research groups.[23] In climate science the research is often not only scattered between different research centres but between different generations of scientists.[24] No modern climate model is built directly from scratch. Climate scientist frequently rely on bits of coding or even whole model parts that were originally developed for a predecessor of the current model, often by a previous generation of scientists (Knutti et al., 2013). Thus, decisions about how to model specific processes made 'back in the days' are still present in today's models. How these past decisions influence the performance of models can be hidden in many ways within the models (Winsberg 2012). Further it is not, for the so inclined scientists, "foreseeable how methodological choices in model development will shape modeling results in the long run" (Parker and Winsberg, 2018, p. 141; for an example see Lenhard, 2018, pp. 839–840).

On top of the intricate model history, other features of climate modelling, such as the high interdependency between different model parts, the fuzzy modularity and the need for tuning, mean that ultimately the model can only

[22] This argument was first made by Winsberg (2012), see also Parker and Winsberg (2017) and Winsberg (2018, pp. 130–153).

[23] Besides climate science, examples for this kind of fractured research can be found in modern physics amongst other research fields. One might, for instance, think of medical research (Kukla, 2012) or modern astrophysics. For an example from the field of gravitational waves, as made by Collins (2014), see Chapter 4.2.1 and see also Collins (2017).

[24] Research centres usually develop their own models although cooperation between different institutes are, of course, taking place. One such example is the new ICON modelling framework that the Max-Planck-Institute for Meteorology developed together with the German Weather Service (Zängl et al., 2015). Furthermore, models might also incorporate specific parts that were originally developed for other models (Parker, 2018).

be fully evaluated within the context of the complete model. It is, therefore, often difficult to say whether adding a new element to the model delivers the desired results because of the new component on its own or is not also the outcome of some compensating effect resulting from the interdependency of different elements of the model. This "epistemic opacity" or "analytic impenetrability" (Baumberger et al., 2017; Humphrey, 2004; Lenhard and Winsberg, 2010) of climate models also cannot (at least practically) be resolved by reconstructing the model completely from the ground up, as Winsberg points out:

> Of course the modeller could – in principle – rework the entire code. The point is, however, that in even moderately complex cases, this is not a viable option for practical reasons. At best, this would be far too tedious and time-consuming. At worst, we would not even know how to proceed. (Winsberg, 2018, p. 143)

In this context it is evidential why any claim similar to that of Jeffrey's that the scientist, instead of making value judgements, should factor the uncertainty estimates concerning different methodological options into their overall assessment cannot be maintained in the context of complex computer simulations. As Winsberg (2012) notes, climate scientists (and scientists more broadly (Steel, 2015)) cannot be viewed as perfect Bayesian actors.[25] The climate system and the models that scientists work with are too complex for any one person to have a complete understanding of the effect of all trade-offs and prioritising on the model.[26] Put in a different way, the high complexity of climate models with its hundred thousands of lines of code, decade long construction history,

[25] Winsberg (2018, 2012) points out that, contrary to Douglas' (2000) claim that the need for inductive risk assessment in the context of methodological decision making, shows the inevitably of social value judgements in science, this does not, in and of itself, refute the assertion that scientist cannot avoid social value judgements. One might still claim that the scientists only have to factor all issues with a particular methodological choice into a probability assessment, which they can pass on as an expert judgement to stakeholders, provided that we assume a classical Bayesian understanding of probabilities and the actors holding them (see Chapter 3.1.2). Winsberg discusses this as the "Bayesian response to Rudnerian and Douglasian arguments from inductive risk" (BRAIR). For a more formal discussion of the Bayesian argument, see Steel (2015) and why it fails in the context of social values in climate science, see Parker & Winsberg (2018) and Winsberg (2018).

[26] In practice, a insufficient documentation of different modelling steps can make the models even more obscure (e.g., tuning, see Chapter 2.1).

its "fuzzy modularity" (Lenhard and Winsberg, 2010) and ever further development make it hard to imagine where scientists should even begin to disclose the whole reasoning process behind every decision ever made in the model construction process (Winsberg, 2018, p. 143). The labyrinthine way of the "nooks and crannies" (Winsberg, 2012, p. 132) of climate modelling makes it impossible to fully trace the consequences of all possible social value judgements made in the model building process.

To be clear, the crux here is not that every epistemic gap is necessarily filed by social values but that, when reviewing a whole climate model, it is not possible to evaluate retroactively to what extent and at what point exactly what kind of social values were a relevant factor in a decision-making process, as well as in what way and if at all they have influenced today's model outputs.

3.1.3.3 Coarser uncertainty quantification and other possible counterarguments

Originally, the discussion about the influence of social values in climate science mostly centred on uncertainty quantification derived from MIPs. In this context there has specifically been some disagreement about how severe the influence of social values actually is. While Winsberg (2018, 2012, 2010) has argued for the possibility of social values filling the void left by epistemically unforced decisions in climate modelling, which cannot be (fully) accounted for during the evaluation process of models, others (Parker, 2014; Schmidt and Sherwood, 2015) have argued that this problem would be at least significantly reduced when scientists are not forced to give a precise estimates of uncertainties but rather are given the option to express uncertainty in ranges of probability. Such a more coarsely grained scheme for uncertainty quantification is what is used by the IPCC for scientists to express their degree of certainty in assessing the current state of climate research. In the *Guidance Notes for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties* (Mastrandrea et al., 2010) authors of the IPCC assessment report are given a guideline which they can and should refer to when conveying uncertainties (see Figure 2).[27] Instead of requiring the scientists to express uncertainty estimates in fixed and precise numbers, they are given wider intervals. The issue of social values in climate modelling can be mitigated this way, argues Wendy Parker:

[27] A further discussion of the *Guidance Notes for Lead Authors* will follow in Chapter 3.3.3.3.2.

> even if social values sometimes do come into play in the model development process in the ways suggested by Winsberg, the influence of those values on estimates of uncertainty will be reduced when coarser estimates are given. The influence will be reduced insofar as choices in model development will less often make a difference to the uncertainty estimates produced. (Parker, 2014, p. 28)

Figure 2: Likelihood scale from the Guidance Note for Lead Authors for the 5th assessment report.

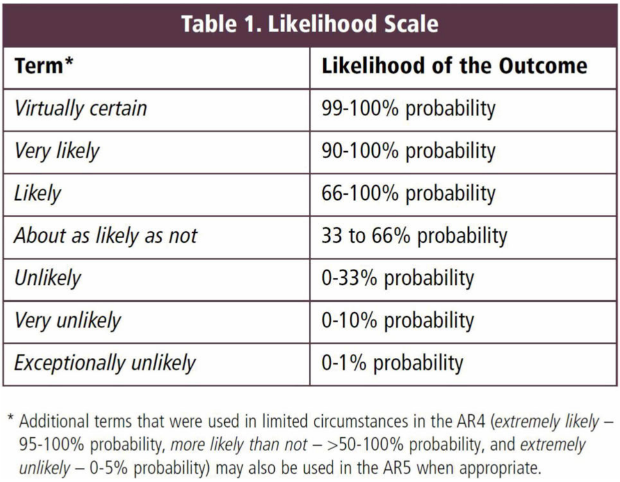


Table 1. Likelihood Scale

Term*	**Likelihood of the Outcome**
Virtually certain	99-100% probability
Very likely	90-100% probability
Likely	66-100% probability
About as likely as not	33 to 66% probability
Unlikely	0-33% probability
Very unlikely	0-10% probability
Exceptionally unlikely	0-1% probability

* Additional terms that were used in limited circumstances in the AR4 (*extremely likely* – 95-100% probability, *more likely than not* – >50-100% probability, and *extremely unlikely* – 0-5% probability) may also be used in the AR5 when appropriate.

Source: Mastrandrea et.al., 2010, p. 3, Table 1

At first glance this might be a satisfying solution, but a closer look reveals that it does not solve the underlying problem. It is likely that a coarser scale for communicating uncertainties will reduce the influence of social values. But the complexity of climate modelling prevents us from evaluating how well this actually works. It is for the same reason that it is not feasible for scientists to keep track of every possible way in which values might have an influence on the model-developing process and then communicate resulting uncertainties to policymakers and the public. The number of epistemically not fully determined decisions are so large and stretched over such a big timeframe that it seems unimaginable how it should be possible to monitor or quantify the decision-making process in all its details. Nor would it be a viable solution to advice scientists to go on to develop *every possible* variety of a model and then evaluate the impact of every single decision made on the performance of the model with statistical methods. As it is usually a multiple year-long project to build a climate model of the scale of an ESM, it is practically not possible to build every thinkable versions of a model. Climate scientists do use *model intercomparison projects* (MIPs) to explore structural uncertainties, but MIPs are not statistical evaluation methods in the sense of, for instance, a Monte-Carlo study (see Chapter 3.3.3.3). They do not sample from the whole space of all models. Instead, they are better described as "ensembles of opportunities"[28] (e.g., Parker, 2010, pp. 270). Parker and Winsberg note in this context that climate models are used precisely "because it can be very difficult to reason about such systems without them" (2018, p. 140). That is, when reasoning about the effects of some modelling decision it often requires scientists to make use of their previous experience with the available models. However this makes it questionable "how 'escapable' the model-based influence of nonepistemic values" (Winsberg and Parker, 2018, p. 140) is even if one resorts to giving coarser estimates.

What is more, the option of a more coarsely grained uncertainty scales brings with it its own new sort of value judgements. Scientists now have to make a decision which kind of scale to use. And the underlying assumptions might well be influenced by social values, as Winsberg argues:

> it seems clear that at least sometimes it is a consideration of the likely applications of an uncertainty report that guide the choice between a wider and more confident report and a narrower and somewhat less confident report. Perhaps a narrower, even somewhat less confident interval is thought to be more useful for policy makers. In such cases, social values are once again playing a role. (Winsberg, 2018, p. 149)

On the other hand, one might also question whether the gaps left by methodologically unforced decisions are necessarily filled by social values. Parker argues that oftentimes pragmatic considerations instead of social values are the decisive factor:

> Suppose a group of climate scientists is further developing their climate model now that more computing power is available. Which physical process

[28] MIPs will be further discussed in Chapter 3.3.3.3.

> should they "add" to their model next? Suppose their choice is epistemically unforced, i.e. they cannot argue on purely epistemic grounds that one process in particular should be added next. Must their choice then either be arbitrary or determined by social values? No. Pragmatic factors can also fill the gap. For instance, the scientists might already have in hand some computer code for process P but not for processes Q, R, or S. Or they might judge that it will be much easier to incorporate P than to incorporate Q or R or S, given past choices in model building. Or they might be experts on P but have much less understanding of Q and R and S. Or it might be that a leading modeling group incorporated P for reasons like those just identified, and now it is seen as de rigueur for state-of-the-art climate models to include P. And so on. Indeed, it is plausible that pragmatic factors like these often influence or even determine model development choices. (Parker, 2014, p. 27)

It is quite plausible that pragmatic considerations can be significant in climate model building. But is that sufficient to completely rule out the potential influence of social values in climate modelling? Clearly not.

First of all, one might ask what "pragmatic" in this context even means. What exactly differentiates social from pragmatic values? And are social and pragmatic necessarily mutually exclusive? Do not pragmatic and social considerations sometimes overlap? One might very well imagine a situation where the decision to fall back on a pragmatic choice such as relying on a model part that is already well known to some of the scientists involved might be influenced by underlying social values. Those may be that a timely solution is valued more even, if it is at the expense of a possibly more precise or detailed answer, because climate change is an issue requiring urgent answers from science. More generally speaking, the decision to go with a pragmatic solution is always based on the (possibly social) value-laden decisions that a pragmatic approach is justified in this situation, as Anna Leuschner rightly notes (2016, p. 79). Furthermore, Leuschner argues, even when a certain piece of code is chosen on the basis of social-value-free pragmatic factors, it is still quite possible that the development of that model component or even just a part of the model component has been influenced by non-epistemic assumptions at one point or another.

Here again it becomes apparent how deep within the "nooks and crannies" (Winsberg, 2012, p. 130) of complex computer modelling value judgements can lie. Pragmatic factors or not, the questions remain to what extent it is actually possible to mitigate the influence of social values and how one would even go about measuring this. Retroactively, there is neither a way of knowing for certain how much influence social values had on a climate model nor is it in any way possible to keep track of every thought that went into the construction of a model. What is so significant is not the particular impact of social values in climate modelling but that we have no way of fully tracing their influence.

But if we cannot retrace and evaluate the influence of social values, it seems also not feasible to restrict exactly what kind of social values should be allowed to impact the scientific process. Kirsten Intemann (2015) proposes that social values can be considered adequate in climate modelling as long as they are democratically supported beyond science. But it is not clear how this could actually be put into practice, let alone be monitored, even if one considers a less complex field of science than climate science. It should be noted that Intemann does not expect scientists to consult the general public at every step of the way. She argues that allowances should be made for a certain flexibility how stakeholders are chosen and to what extent their values actually should match those of whom they represent so that "modeling decisions can be more or less justified in degrees depending on the extent to which social and epistemological aims are clear and there is evidence that they would be broadly endorsed" (Intemann, 2015, p. 228). Further, she argues that the scientist is not required to consult the stakeholders at every step of the way, rather scientist and stakeholder are in "a process of interactive feedback loops" (Intemann, 2015, p. 288). Still, there is no way of knowing if just democratically determined values (even if only to varying degrees) are really the only ones that play a role in the decision-making process in climate modelling. On the contrary, it seems very questionable to me that they are all necessarily democratically supported. The sheer number of methodological underdetermined decisions, again, makes that very unlikely. After all, scientists are not a homogenous group, so we can expect them to have a variety of values and, as epistemically not fully constrained choices are a consistent feature of climate modelling, it is questionable whether they actually check regularly if their choices are in accordance with the wider societies' values. But whatever might determine those "thousands of methodologically unforced decisions" (Winsberg, 2012, p. 130), we might never entirely know for at least a significantly large number of them.

All this might lead one to question why some philosophers and even climate scientists[29] are so determined to show that the influence of social values in climate modelling is either negligible or can somehow be (democratically) legitimised. I think Winsberg is right when he argues that at the root of this is a misunderstanding and that value-ladenness "is not at all the same as the claim that scientific conclusions are reached in a way that is systematically biased" (Winsberg, 2018, p. 150). Mere value-ladenness is not does not mean outright bias.

3.1.3.4 Systematic bias and wishful thinking

This leads us back to the main concern of the proponents of the value-free ideal: social values in science will inevitably lead to bias and wishful thinking. At first glance, this might seem to be a reasonable concern. After all, climate science seems to be saturated with all sorts of possible social-value-type assumptions. And as climate change is a highly political topic, one might assume that this could give scientist plenty of opportunities to influence the models consciously or unconsciously in a way that suits them best. However, a closer look at the situation shows that this concern is unwarranted.

First of all, the number of scientists involved in the development of a global climate model (at least in the case of those of the complexity of an AOGCM or ESM) make it highly unlikely for it to be possible for one scientist to single-handedly 'sabotage' a model, at least not without being noticed by their colleagues. However, a follow-up claim might then be that climate scientist collectively consciously or unconsciously influence the model developing process in a way so the models are in accordance with their personal social convictions. In the first case, we would imply that there is a grand conspiracy at play involving huge parts of the climate science community. If we disregard this rather outlandish assumption, the second possibility is a little bit more complicate to refute. This claim relies on the assumption that scientists are a monolithic group in one way or another. Looking back at the history of public perception of climate science it has not been an uncommon occurrence for climate-change deniers to accuse scientists of being biased. Either because they (unconsciously) fear that their work would otherwise be redundant or because they cannot separate their work and their personal political convictions, as described in the introduction to this book.

[29] For examples of the former, see also Betz (2013) and for the latter, see Schmidt and Sherwood (2015).

The worry that the value-laden background assumptions of scientists can have an inappropriate effect on scientific research in general is also voiced by many feminist philosophers of science. Longino (2002, 1990), for instance, advocates for a pluralism of perspectives in science. The hope is not that, within a diverse group of scientists with different backgrounds (social, racial, political, and so forth) compared to a group of scientists with a homogeneous set of values, the influence of social values can be prevented but that their impact will be easier to detect. Leuschner (2012a) argues that the IPCC's structure, regarding the selection of scientists, is similar to such a kind of pluralism.[30] As an intergovernmental organisation the IPPC specifically selects the authors "taking into account the range of scientific, technical and socio-economic views and backgrounds, as well as geographical and gender balance" (IPCC, 2023). This, argues Leuschner (2012b, pp. 176–177), fulfils two purposes: on the one hand, the hope is that all involved countries feel included and are, therefore, more inclined to implement mitigating climate policies. But on the other hand, there is also the epistemic expectation that this will ensure that all relevant knowledge and data is taken into account under the assumption that scientist have unique and specific scientific, cultural and political knowledge about the region or country they come from. The IPCC itself also states that the reason for diversifying the field of participating scientists is "to ensure that reports are not biased towards the perspective of any one country or group of countries and that questions of importance to particular regions are not overlooked" (IPCC, 2023). The purposeful inclusion of minorities in science is not an unscientific act fuelled by social values but rather follows good epistemic consideration.

[30] To be more precise: Leuschner agrees with Longino on the necessity of a pluralist approach to science in order to reveal hidden value assumptions, but she also criticizes that Longino's idea how this should be implemented in practice "suffers from an inherent circularity" (Leuschner, 2012a, p. 197). Longino argues for a pluralism that simultaneously demands that everybody ought to be able to participate in the critique of scientific discourse but also to exclude any unqualified opinions. This is contradictory and circular, notes Leuschner, as it is not clear how any kind of standards defining what qualifies contributors are is to be determined, without constricting the pluralistic process of including as many perspectives as possible. Inspired by Kitcher's concept of "deliberators" (2001), Leuschner argues for a pragmatic and situation-specific solution for this problem including pluralistically organised but politically installed expert groups, which would evaluate scientific practices and findings, such as it is the case with the IPCC.

However, it is not just the outer socio-scientific structure of organisations like the IPCC that implement procedures to constrain bias, but also the complexity of both the climate system and the models in and of itself that helps prevent wishful thinking affecting research. Complex climate models of the scale of an ESM or AOGCM are used for multiple purposes. Contrary to the public perception, the main goal of climate science is not just to further pin down exactly how the climate will change under a certain emissions scenario but rather advance the understanding of specific climate processes and the climate system as a whole. Considering the amount of money and effort going into developing a new model of this type, it seems obvious that scientists and financial backers are very much interested in developing models that can be used for a variety of purposes.[31] However, this also means that different researchers or research groups that are involved in the development process come with slightly different agendas to the table.

What is more, the argument that climate scientists could just adjust the models to their own preference loses in strength when one considers that various parameters of interest do not exist in isolation in the models. Quite often changing one parameter also directly influences other. Thus, as not all variables and processes that the scientists would like to explore with the model can be equally well represented, scientists are again confronted with having to make trade-offs with respect to their competing preferences. Here again is a kind of pluralism at play.[32]

Climate science, thus, gives us a perfect example why possible value-laden deliberations within the inner-scientific process are by far not as much of a threat to science as the discussion about them in philosophy of science lets us believe. Not because they are so rare or obvious to spot. On the contrary, there is a myriad of ways in which social values might influence scientific processes, yet because of that they are not just unavoidable but also mostly epistemically harmless. Additionally, as this chapter has shown, fuzzy modular-

[31] There is currently even a trend to develop models that can simultaneously be used for climate modelling purposes as well as weather prediction. This was first done with the Unified Model of the Met Office in the UK. In Germany, the Max-Planck-Institute for Meteorology and the Deutscher Wetterdienst (German Weather Service) have also joint forces to develop a shared model framework (ICON).

[32] Besides that, one has to remember that climate scientists rely on more than ESM to assess the impact of climate change, such as a variety of global models, regional models and empirical data from observation and experiments from different fields of research and expert judgement (see Chapter 3.3.3.4).

ity, compensating effects and entailing trade-offs mean that it is simply impossible to tune the 'perfect' model. The complexity of the climate modelling is simultaneously the reason why value-laden considerations may enter the climate model building process as well as the very feature that protects climate science from unwanted (conscious or unconscious) political or social influence. The consequences of every modelling decision can be so manifold and inscrutable that it makes it in fact much harder to argue that it is even possible for scientists to influence the models effectively in a way that suits their own social or political beliefs (Parker and Winsberg, 2018). Further, representational risks, which are likely a bigger source of social value-laden decision making in climate modelling than inductive risks do not constitute "influences that make it more likely that one conclusion rather than another will be reached" (Winsberg, 2018, pp. 150).

3.1.4 Conclusion

Nowadays the vast majority of philosophers of science accept that value judgements are an unavoidable element of science. But there is a lively discussion ongoing about what exactly the appropriate role of social values is and how the non-epistemic, social realm can be constricted. While Intemann, for instance, argues that "value judgments are legitimate when they promote democratically endorsed epistemological and social aims of research" (2015, p. 217), Douglas (2009) suggests that social values at stages internal to science should be restricted to an indirect role. Both positions have in common that they argue that the influence of social values must be restricted and limited to specific cases one way or another.

When it comes to a differentiation between the direct and indirect role of values, it seems questionable if that distinction is particularly helpful here. The distinction that Douglas makes is fuzzy to begin with. Douglas notes that there are exemptions for both kinds of roles of values. Thus, values might play a direct role in the inner-scientific process such as when the ethical implications of methodologies have to be unexpectedly reassessed. In the same way, according to Douglas, social values might also be inappropriate in the pre-scientific context when they undermine the core scientific goal of gaining knowledge.

Further, inductive-risks assessment is not the only way in which values play an unavoidable role in science-internal processes. As Harvard and Winsberg (2022) note, when it comes to representational risks, determining the appropriate role for social values by distinguishing between a direct and indirect role of social values is no longer a viable route.[33] Predictive preferences and cost deliberations interfere in a much more direct form. Note that these kinds of value judgements are not just decisions under uncertainty but, in fact, decisions of scientific research objectives.

It has to be emphasised here that predictive preferences are not necessarily priorities which are set before the model is developed. There might be certain priorities that scientists define before setting out to create a model. But these have to be rather general as the model cannot be planned in all its details from the beginning. Climate modelling involves a certain degree of tinkering and trying out which method works best for the specific model (Held, 2005).[34] The complexity of these kinds of models makes it impossible to anticipate every decision necessary in its construction. Thus, the research goals and priorities are constantly under some threat of having to be reset and readjusted (to some degree at least). Further, the number of people involved, often over several generations, makes it unlikely that they all share and abide by the same interpretation of these priorities. This is further complicated by the fact that by relying on model parts, whether whole parametrisations schemes or bits of code originally developed for different models (in other words by not constructing the model from scratch) choices made decades ago will constrict the modelling process. Therefore, one cannot simply view these *predictive preferences* as pre-scientific goal setting, that is the kind of social-value interference that even most proponents of the value-free ideal see as unproblematic (Douglas, 2009, p. 45).

The value-free ideal has always been an illusion, created by science to protect itself against unwanted interference from religion (Rudner, 1953) or for fear of losing its authority (Douglas, 2009, p. 79). As much as the supporters of the value-free ideal have tried to deny it, values have always been part of science. Science is neither all of a sudden overrun by social values nor has science become unreliable and biased. Such a view of science overlooks the fact that this is a state that science has always been in. The increase of complex systems as the subject of research in science just makes the illusion of the value-freeness of science that much more obvious.

[33] Although the authors come to the conclusion that where the risk of endorsing as false fact is concerned, limiting social value judgments to an indirect role at an internal scientific stage might still be a good way to rule out wishful thinking (Harvard and Winsberg, 2022).

[34] See also Chapter 4.2.2.

It is, of course, quite understandable why some climate scientists have initially reacted hostile to the suggestion that social values might be at play in climate modelling (e.g., Schmidt and Sherwood, 2015).[35] After all, they are regularly under fire from climate-change deniers, who accuse climate scientists of being misled by their own personal convictions. As Proctor puts it:

> Value-freedom is an ideology of science under siege – a defensive reaction to threats to the autonomy of science from political tyrants, religious zealots, secular moralists, government bureaucrats, methodological imperialists, or industrial pragmatists asking that science be servile or righteous or politically correct or practical or profitable. (Proctor, 1991, p. 68)

Climate science has very much been such a "science under siege". In such a situation, it might be tempting (and often initially successful) to insist on the value-freeness of one's own research, but scientists do science as a whole a disservice when they keep insisting on practicing value-free science. Asserting that the value-free ideal is still upheld – born out of an instance of self-defence – chances are high that it will backfire in the long run because the more the complexity of the systems that scientists investigate increases, the less likely it will be that scientists can successfully hide behind an apparent value-freeness. Science – as most human enterprises – cannot, has never and will never be value-free.

While there might have been originally some scepticism from within the climate-science community when philosophers began to discuss the role of social values in climate modelling, it also has to be noted that the role of social values in climate science are actually openly discussed in the latest IPCC report. Particularly, predictive preferences (though not named as such) being an unavoidable element of climate modelling are highlighted:

> Social values are implicit in many choices made during the construction, assessment and communication of climate science information (Heymann et al., 2017a; Skelton et al., 2017). Some climate science questions are prioritized for investigation, or given a specific framing or context, because of their

[35] Gundersen has shown in a small study of a group of Norwegian climate scientists that many of them stand by the value-free ideal, although the scientists also note that it is sometimes difficult to guarantee the value-freeness in practice. Gundersen remarks as well that some scientists observe that "strict adherence to the value-free ideal can undermine policymakers' perception of the relevance of experts' opinions" (2020, p. 113) when it makes them to be too cautious in conveying the significance of their findings.

> relevance to climate policy and governance. One example is the question of how the effects of a 1.5°C global warming would differ from those of a 2°C warming. [...] Likewise, particular metrics are sometimes prioritized in climate model improvement efforts because of their practical relevance for specific economic sectors or stakeholders. [...] Sectors or groups whose interests do not influence research and modelling priorities may thus receive less information in support of their climate-related decisions. (Chen et al., 2021, p. 172)

This quote also points us toward another important distinction that needs to be made. I have argued above that value judgements in climate science are in most cases epistemically harmless; however, that does not mean that there are not at the same time some non-epistemic risks that can arise out of predictive preferences, to the extent that, e.g., a lack of attention towards the predictive preferences of underprivileged community can cause harm when their need for a particular kind of knowledge is not taken into account (Harvard and Winsberg, 2022; Parker and Winsberg, 2018). One way to mitigate this risk seems to be (again) a pluralistically organised scientific community (Jebeile and Crucifix, 2021).

So far the discussion of social values in science has often centred on the argument from inductive-risks. But, as has been shown in case of climate science, representational risks are also an unavoidable part of developing a complex computer simulation. Further, we have also seen that with the increasing complexity of science it becomes more and more impossible to retroactively make those value decisions explicit. As I have argued above, this should not be seen as an epistemic problem. Instead value judgements ought to be regarded as a necessary part of science. Particularly when it comes to complex computer simulations, they fill gaps left by epistemic and methodological underdetermination. While the complexity of the system introduces an inability to trace the effects of social-value deliberations through the model building processes, it also works simultaneously as a safeguard against the directed value-laden manipulation of the models or, in other words, the complexity 'inoculates' the models against wrongful influence of this kind. On the one hand, the specific values are so numerous and diverse and quite often do not even have the "right form" (Winsberg, 2018, p. 151) to make the models biased in a specific way. On the other hand, the number of scientists involved works as an insulation and corrective tool against individual bad work. Particularly, when the group of modellers is sufficiently diverse, it makes it more likely that "later choices in model development can 'undo' the effects of earlier ones" (Parker and Winsberg, 2018, p. 141, see also Jebeile and Crucifix, 2021). Whatever may be the case, the issue is not whether or not the scientist's decisions are directly guided by social consideration, but that we can no longer retroactively tell whether that has been the case or not (Winsberg, 2012).

Under these circumstances it is also questionable if a distinction between cognitive and social values is still useful to determine the appropriate role for value judgments in science. Even if we were certain that all modelling decisions were determined by cognitive values alone, one still cannot be sure that these are not affected by the specific social context under which they were constructed. Considering Longino's (2008) claim that we can just as well imagine alternative set of cognitive values, which are as well justified as the traditional ones, one has to at least question if in a similar way specific modelling cultures might be in a hidden way informed by social or ethical background assumptions (see Chapter 3.1.1.2). So it does not even make sense in this context to discriminate between some kind of science-internal, appropriate and extra-scientific values that are only under very specific circumstances allowed to interfere with scientific processes.

Though the fear of the influence of social values on science is historically understandable, we need to change our perspective on values in science. Against the backdrop of the vast number of epistemically not fully constrained decisions and the new epistemic challenges of dealing with highly complex systems, the discussion about values in science must shift from a discussion of what the appropriate role of values in science is to what an inappropriate role would be. Douglas has argued that "values should never suppress evidence, or cause the outright rejection (or acceptance) of a view regardless of evidence" (2009, p. 113). Despite the vast variety of necessary roles that values judgements can assume in the construction and evaluation of complex computer simulations, this also seems to be a prudent approach for inductive risks. However, predictive preferences, which seem to be the biggest source of possible social value-laden assumptions, do not hold a clear risk for scientists to outright disregard evidence as it is primarily a question of research objectives (see Harvard and Winsberg, 2022). Here the bigger, non-epistemic concern is that some underrepresented stakeholders might get less information concerning their particular circumstances. Nevertheless, as has been argued above, from a purely epistemic perspective, the risk of inappropriately influencing climate-model construction through wishful thinking or deliberate bias does not seem to be a particular high risk.[36]

3.2 Model, theory and observation

3.2.1 Introduction: from handmaiden to a life of their own

For a long time the experimental and practical part of science has been somewhat neglected by philosophy of science. Until well into the middle of the 20th century the discourse in philosophy of science has focused primarily on a discussion of scientific theories. While Francis Bacon saw the experiment at the centre of the scientific enterprise, by the time philosophy of science had become a discipline of philosophy in its own rights, at the beginning of the 20th century, observations and experiments had been cast into the role of the "handmaiden of theory" (Gooding, 2000, p. 119). Meaning that the primary purpose of the empirical part of science was seen as to provide data to evaluate theories. According to this theory-focussed view of science, experiments and observations are only of relevance to science once a theory has been developed and needs to be tested. They were considered to be of little philosophical interest on their own.

This disregard of the experimental and observational element of science in the history of philosophy of science becomes most apparent in the approach that logical empiricism and critical rationalism take to this issue. While the logical empiricists focused on the logical and theoretical foundation of science, they reduced the empirical part of science to producing basic observational sentences (*Beobachtungssätze*).

This indifference towards the practical part of science in the first half of the 20th century did not just hold for logical empiricism. Karl Popper, who in

[36] In effect, cases where climate-change deniers have argued against mainstream science qualify as exactly one of those situations where the deniers have neglected the evidence in favour of personal beliefs. Naomi Oreskes and Eric M. Conway (2010) show compellingly how a small subset of scientists, often paid by specific interest-groups, such as oil companies, have disputed a variety of scientific claims from smoking causing cancer to climate change over many decades because of their dislike of governmental regulations.

many ways opposed the philosophy of the logical empiricism, still saw the experimenter as someone who does the legwork while being fully guided by the theoretician:

> The theoretician puts certain definite questions to the experimenter, and the latter, by his experiments, tries to elicit a decisive answer to the questions, and to no others. […] But it is a mistake to suppose that the experimenter proceeds in this way 'in order to lighten the task of the theoretician', or perhaps in order to furnish the theoretician with a basic for inductive generalizations. On the contrary, the theoretician must long before have done his work, or at least what is the most important part of his work: he must have formulated his question as sharply as possible. Thus it is he who shows the experimenter the way. But even the experimenter is not in the main engaged in making exact observation; his work, too, is largely of theoretical kind. Theory dominates the experimental work from its initial planning up to the finishing touches in the laboratory. (Popper, [1935] 1959, p. 107)

Popper does not consider the work of the experimenters as completely superfluous, but he also does not see them as the ones who take initiative. The experimenter, according to Popper, is not in a position where they could contribute anything substantially new to science on their own. On the contrary, the job of the experimenter is seen here as only to confirm or falsify hypotheses as instructed by the theoretician. Thus, to Popper the experiment takes a subordinate role to the theory. It provides evidence but cannot in itself provide new scientific insight. It is for this reason that Popper gives the *theory* that much more attention than the *experiment* in his writings.

Just as in the case of the rise of the value-free ideal this development coincides with the rising popularity of the distinction of the co*ntext of justification* from the *context of discovery* made by Reichenbach (1938). As already discussed in more detail in Chapter 2.2, this distinction became very popular among philosophers of science of the 20th century, when it came to separating the realm of philosophy of science from that of psychology and social studies. Proponents of this view are of the opinion that philosophy of science should focus on logical justification of a scientific discovery, not the practical path leading to it. The question of how a scientific fact, theory or law is discovered is thereby made a matter of sociology or psychology but not philosophy. In that respect the DJ distinction also played an important role in directing the attention of philosophers of science towards theories (Schickore and Steinle, 2006a). Experiments and observations were seen to be only tangentially philosophically important insofar as experiments are the way to provide empirical evidence for theories. However, the process through which the necessary data for this is acquired was seen as by and large not appealing to philosophical contemplations. The experimental part of science was, thus, (dis)regarded as (for the most part) an element of the context of discovery and, thereby, cast aside as philosophically uninteresting.

For most of the 20th century the default position seemed to be to omit the experimental part of the scientific enterprise from the philosophical discourse, or as Gooding put it: "Experiment seems to be an epistemological football – essential to the game, but of no intrinsic philosophical interest" (Gooding, 2000, p. 122). However, beginning in the late 1970s, philosophers and sociologists of science increasingly started to question the theory-dominant view of science and set out to bring "studying scientific practice, what scientists actually do" (Pickering, 1992, p. 2) back into the limelight. During what is today sometimes described as the *practical turn* or the *new experimentalism* they began to explore the different functions and characteristics of experiments and other aspects of actual scientific practice (e.g., Ian Hacking, David C. Gooding, Allan Franklin, Nancy Cartwright). Though they all highlighted the necessity to include experimental scientific practice in philosophical discussion about science, there is a certain disagreement to what extent experiments can be considered independent from theory (Feest and Steinle, 2016; Gooding, 2000).

Ian Hacking for instance argues that experimentation can happen independently from theories. By means of a number of examples from the history of physics, Hacking shows in his well-known book *Representing and Intervening* how varied the dynamic between theory and experiment in actual scientific practice can be (1983, pp. 149–165). While sometimes the theory preceded the experiment, quite often experiments were done independently of a specific theory. One example Hacking gives of such a case is the early days of optics where experiments were done without any fully established theory. Another case concerns the discovery of cosmic background radiation which was discovered experimentally independently of a corresponding theory that was developed elsewhere at the same time by different scientists. For Hacking experiments are as philosophically intriguing and important in knowledge generation as theories. Further, the experimentalist does not rely on the theorist to provide them with a hypothesis to (dis-)confirm.[37] The experimental part of science has, as Hacking puts it, "many lives of its own" (1983, p. 165).

Despite these more recent trends, there is still quantitatively more literature in philosophy of science about the theoretical part of science than about the practical, empirical part. Something which can also be observed in the context of philosophy of climate science. While much has been written and said about the use of computer simulation in climate science, less can be said so about the creation and evaluation of observational data.[38]

In the following I will argue that, in fact, the process of empirical observation making in climate science is just as philosophically interesting as the computer simulations used to model the climate system. Furthermore, a closer look at the complex process of the production of climate data and the intricate relationship between observational data and climate models will reveal that conventional ideals about the role of observations in science, similar to those expressed in the quote by Popper above, cannot be maintained. As will be shown, making observations and constructing models are neither fully separated processes nor can it be said that observations provide irrefutable benchmarks to distinguish good models from bad ones. Particularly, a widely discussed controversy about satellite data will show how a widespread presence of these idealisations of scientific procedures in the public's understanding of science can be capitalised on by climate-change sceptics and so inclined interest groups to sow doubt about the trustworthiness of climate science.

I will begin in Chapter 3.2.2 with a short general philosophical debate of some relevant philosophical concepts, particularly *theory-ladenness* and *models of data,* before I turn to the specific case of observational data in climate science. But before doing so, it seems prudent to first consider the definition of the term *observation*.

[37] All this has, of course, to be understood in the context of Hacking's *entity realism*, according to which experiments can confirm the existence of entities independently from theories. As he famously puts it in respect to the use of electrons in experiments: "if you can spray them, they are real" (Hacking, 1983, p. 24).

[38] Some noteworthy exemptions are Edwards (2010); Guillemot (2017, 2010); Lloyd (2012); Parker (2020, 2017). Edwards (2010) specifically writes an extensive historical account of the development of a meteorological and climatological infrastructure.

3.2.1.2 Observation

In the context of the empirical part of climate science, scientists usually describe their work as making and processing *observations*. But what actually defines *observations* and what makes them distinct from *experiments*? And to what extent can this distinction be made at all? To answer this question let us first take a look at a short history of the term, as recounted by Daston (2011) and Daston and Lunbeck (2011).

From a historian's perspective, the use and meaning of the terms *observation* and *experiment* in science itself has changed more than once over the last four centuries from synonyms to antonyms, as Daston explains:

> In the period from the early seventeenth to the mid-nineteenth century, the relationship between observation and experiment shifted not once, but several times: from rough synonyms, as in the phrase "observations and experiments" that had become current in the early seventeenth century, to complementary and interlocking parts of a single method of inquiry throughout much of the eighteenth and early nineteenth century, to distinct procedures opposed as "passive observation" and "active experiment" by the mid-nineteenth century. (Daston, 2011, p. 82)

While empirical science as such becomes gradually more relevant in the late 17th century, the term *experiment* becomes narrower and now refers to "deliberate manipulation" (Daston, 2011, p. 85) or what Bacon called *artificial experiment*; the term *observation* becomes wider. Observation making, disregarded by medieval scholars "with conjecture because its results were uncertain" (Daston, 2011, p. 104), had become an activity so relevant to science by the middle of the 18th century that it had become "a way of life" for many in the scientific community, dictating their daily routine (Daston, 2011, pp. 101–104).

Towards the end of the century, notes Daston (2011), observations had become a full-fledged 'instrument' of thinking and reasoning in its own right, including repetition of observations and comparison to others. Observation making as a whole had taken a distinct, methodological, systematic and communal character. It had become something which most scientists saw central to their work and reasoning processes.

In the 19th century the terminology underwent a new shift, which singled out the experiment as the activity requiring real talent and training, contrary to observations:

> starting in the 1820s, prominent scientific writers began to oppose observations to experiment, and to vaunt the prestige of the latter over the former. In this new scheme of things, experiment was active and observation was passive: whereas experiment demanded ideas and ingenuity of the part of a creative researcher, observation was reconceived as the mere registration of data, which could, some claimed, be safely left to untrained assistants. (Daston and Lunbeck, 2011, p. 3)

Seeing observation making as a passive exercise that could also be done by the untrained was not a degradation of the relevance of observation to science but interpreted as an advantage to science. By being able to outsource observation making to 'untrained forces', scientists hoped to make sure that the data remained 'objective' and unspoiled by the scientist's theories. This attitude is also reflected in 20th century philosophy of science, according to Daston and Lunbeck. The attempt of the philosophers of logical empiricism to create a scientific system in which any theory can be retraced to observational protocol sentences "would render observation in a language as close as possible to the raw data of perception" (Daston and Lunbeck, 2011, p. 5).

These days two different but at the same time overlapping definitions of the two terms are very common. Today scientists usually loosely ascribe the term *observation* to data collection in a fixed target system, whereas experiments in science traditionally include manipulations of the target system. Both can be understood as empirical, scientific practices from different ends of a spectrum. While experimentation demands active intervention by the scientists, usually in a laboratory setting, making observations is seen as a much more passive activity requiring often the skilful application of measuring instruments but no interference with nature.

There is also another use of the term *observation*, that is not as widespread in science but highly prevalent in philosophy of science. This definition also sees observation as a passive activity but narrows it down even further. Here *observation* refers to the sheer perception or detection of data, quite often as part of an experiment but also as detection of natural phenomena with or without instruments.[39] This definition of the term will have specific relevance in the context of the following discussion of theory-ladenness of observations.[40]

[39] For the difference in the use of the term observation see, also Shapere (1982).

[40] Hacking incidentally, though dedicating a whole chapter in *Representing and Intervening* to the topic of observations, is rather imprecise in his use of the term *observation*.

One common feature of both definitions is that they both seem to have retained the notion of observation as a passive and experiments as an active undertaking. But as we will see in the following in the case of climate science, disregarding observations as just passive perception overlooks how much active knowledge of the matter at hand and training in the use of instruments are required in scientific observation making.

3.2.2 Theory-ladenness, underdetermination and models of data

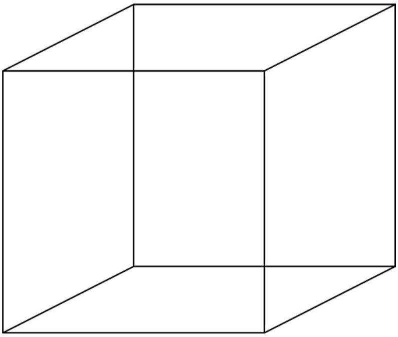
The notion that theories are underdetermined by observations, meaning that there can always be an alternative theory to explain an observation, was first introduced by Pierre Duhem (1906) for theories in physics. W.V.O. Quine (1951) extended this idea to any kind of knowledge claim. Today underdetermination is often understood as the Quine-Duhem problem, as a problem of confirmational holism to the extent that every hypothesis is accompanied by auxiliary hypotheses (Stanford, 2023). That is, when a hypothesis is found not to be in accordance with the empirical observations, the hypothesis cannot be (by logic alone) deemed wrong, as it could also be the case that one of the auxiliary hypotheses is wrong.

Norwood Russell Hanson (1958, pp. 4–30) was the first to introduce the phrase of *theory-laden observations*. Hanson argues that two scientists can observe the same object but 'see' different things. Assuming a proponent of the heliocentric and one of a geocentric worldview watch the sun rise together, he poses the question: do they see different things? Hanson argues that this is in fact the case. When scientists speak of 'seeing' something, they do not refer to the mere process of physiological perception or, as Hanson put it, "seeing is an experience. […] People, not their eyes, see" (1958, p. 6). So looking at the sun is more than the reception of photon particles on the retina of the scientist's eyes. Two

He mainly refers to observation in the context of data detection, but occasionally he gravitates towards a definition that is closer to a definition of observation as experiments with fixed target systems (1983, pp. 155–156, 180). As Malik (2017) points out, by (in effect) also defining experiments as the creation of phenomena in a pure state, Hacking omits great parts of science, which then fall neither into the category of observation, experiment nor theory, such as medical research. Hacking can only maintain this narrow definition of experiments by almost exclusively referring to examples from physics.

people having the same physical premises (e.g., both have equally well functioning human eyes) and observing the same physical object can still 'see' different things, according to Hanson. In this sense, the heliocentrist would look at the sun and see an object at the centre of our solar system, and the geocentrist sees an object circling the earth. How scientists 'see' the world and data they extract from it are, argues Hanson, influenced by their specific background assumptions.

Figure 3: Necker Cube



This, Hanson insists, is not just a kind of interpretation of data: "To interpret is to think, to do something; seeing is an experiential state" (Hanson, 1958, p. 11). The scientist looking at a lab sees various instruments such as specific microscopes or other special instruments. The scientist does not think to interpret the instruments as such, they just see. The layperson, on the contrary, only sees a number of cables and lenses and so forth. "The knowledge is there in the seeing and not an adjunct of it" (Hanson, 1958, p. 22). Hanson compares this to the way we perceive ambiguous images (Hanson, 1958, pp. 8–14). When we look at the Necker Cube (Figure 3), for example, some might see it as from above, others as from below. But we cannot see it in both ways at the same time. Nor would we describe the way we see the cube as an interpretation of twelve specific lines on a paper. We just *see it*; we see *a cube* from above or below, respectively. And someone who has never been taught or who does not have the physical ability to see how twelve lines of the same size, ordered in a specific way, can look like a cube will only see twelve lines. Our prior knowledge guides how and what we see, argues Hanson. It is a one-step process. Seeing and knowing go together, insofar as the way one observes depends on the individual background assumptions and knowledge. Imagine, for instance, a physicist and a child both looking at the drawing of an X-ray tube but which to a child who has not been told otherwise just looks like a scribble of circular and straight lines (Hanson, 1958, pp. 15–19). How we observe the world is context-dependent and observations are theory-laden in the sense that an "[o]bservation of *x* is shaped by knowledge of *x*" (Hanson, 1958, p. 19).

Similar sentiments to Hanson's were also voiced not much later by other philosophers of science like Kuhn (1962) and Feyerabend (1959). It is, therefore, worth mentioning here that, although the concept of theory-ladenness has subsequently been well-established in philosophy of science, not all philosophers follow Hanson's view that all observation is necessarily theory-laden (Feest and Steinle, 2016). Hacking, for example, argues that there are instances in the history of science where scientific discoveries were made without necessarily having the right background assumptions, such as William Herschel's discovery of radiant heat (1983, pp. 167–185).[41] Hacking tells the story how Herschel, after initially noticing how different filters he had used on his telescope transmitted different amounts of heat depending on their colour, began to further experiment with a prism and a thermometer, measuring the heat of rays of light and made further measurements with all sorts of filters. He did, argues Hacking, all of this without having a fixed idea what was actually going on.[42] In the end, Herschel abandoned the experiment. But the reason why he gave up, according to Hacking, was not that he had no satisfying theory but experimental difficulties that he could not overcome. To Hacking being a good observer is much more a question of being skilled at specific observation-making practices than having background assumptions about what is observed, an aspect of observation making that I will come back to in Chapter 4.[43]

[41] Hacking here applies a rather narrow concept of *theory* and defines it as "a word best reserved for some fairly specific body of speculation and propositions with a defined subject matter" (1983, p. 175).

[42] Though Hacking notes that Herschel's first guess of a partially visible spectrum of light coming from the sun was close to our current understanding of what causes radiant heat (1983, pp. 176–177).

[43] Hacking invokes the examples of lab technician without a university degree or the ability of William Herschel's sister Caroline to detect comets (1983, pp. 179–180).

3.2.2.1 Models of data

The idea of *models of data* has been well-established in philosophy of science since it was first introduced by Patrick Suppes (1962). Suppes notes that where the evaluation of theories is concerned one does not simply compare theories to raw observations:

> Theoretical notions are used in the theory which have no direct observable analogue in the experimental data. In addition, it is common for models of a theory to contain continuous functions or infinite sequences although the confirming data are highly discrete and finitistic in character. (Suppes, 1962, p. 253)

Instead, scientists compare *models of theory* to *models of data*. These data models are usually interpreted to be the statistical analysis of the experimental research results.[44] In this sense a model of theory is a specific "realization of the theory" and a model of data is a "possible realization of data" (Suppes, 1962, pp. 252–253).

Two examples will show how this translates into scientific practice. The first example, weather forecasting, which will also give us a first glimpse at how the notion of models of data will be relevant to climate science is given by Baas van Fraassen (2008):

> On the weather forecast website I consult I can find a graph depicting yesterday's temperature plotted against time. This was constructed from data gathered at various stations in the region, at various times during the day – this graph is a smoothed-out summary of the information that emerged from all these data, it is a *data model*. The question about the daytime temperatures in this region of one day ago is answered with a *measurement outcome*, certainly – but that is the graph in question, which is a data model constructed from an analysis of the raw data. (van Fraassen, 2008, p. 166)

Van Fraassen also emphasises that – whether a single measurement outcome or statistically processed models of data are considered – the data is also

[44] Suppes specifies a model of data as "designed to incorporate all the information about the experiment which can be used in statistical tests of the adequacy of the theory" (1962, p. 258). This interpretation of models of data as statistical models can be problematic, as Leonelli (2019) points out, because this excludes certain types of data such as images.

shaped by the circumstances under which the measurements are taken. One has to look at the result as "*this is what the object looks like in this measurement setup*" (van Fraassen, 2008, p. 167).

A similar point of view as far as the role of observation in science is concerned is expressed by Ronald Giere (2006). He argues that every observation, whether done with our own eyes or through instruments, are done from a specific perspective. That is, the use of scientific instruments does not give us a more 'objective' understanding of the world, in the sense that it provides a view from nowhere, free from the personal perspective of the scientists. There might be ways in which scientific instruments can reduce human influence and, thereby, make observation more stable, but they cannot represent the world from a universal perspective, argues Giere:

> The inescapable, even if banal, fact is that scientific instruments and theories are human creations. We simply cannot transcend our human perspective, however much some may aspire to a God's-eye view of the universe. (Giere, 2006, p. 15)

That is, instruments can only provide us with a picture of the world that is taken from a specific point of view.[45] One example of this, Giere (2006, pp. 41–49) provides, are modern telescopes as they are used by astrophysicists. First of all, there is a variety of different kind of telescopes measuring different things: radiotelescopes, gamma ray telescopes, X-ray telescopes, optical telescopes, to name only a few.[46] Furthermore, the actual physical position of the telescope is relevant: e.g., here on earth or in space. All of these telescopes would measure something differently even if we were to point them at the same part of the sky. That means, argues Giere, they show us a particular perspective of the same part of space. On the flip side, this also means that these instruments are also

[45] This, of course, also holds for humans themselves. Humans have, as Giere points out, a specific (trichromatic) colour vision of the world that is the result of the interaction of our body (the perception of light rays on the retina in our eyes) with some physical processes and features of the objects (chemical setup of the object and radiation of light), which not all animals share because they are, e.g., dichromats or tetrachromats (2006, pp. 17–40).

[46] Modern telescopes such as the *Hubble Space Telescope* carry instruments that can measure a wide range of wavelengths, but scientists and technician operating them must still make a decision which wavelengths are relevant for their research questions.

always 'blind' in some respects. Gamma-ray detectors in a telescope are tuned to, well, gamma rays and cannot detect radio waves.

Secondly, the data produced by these telescopes looks at first nothing like the pretty, colourful images we know from magazines or sci-fi films. Before they are useful to scientific research, this data also has to be transmitted, filtered, evaluated and corrected for background noise and measuring errors among other things.[47] All this is done by relying heavily on theoretical background assumptions. The *Hubble Space Telescope*, for instance, produces images through gravitational lensing. That means that scientists rely on the assumption that, according to the theory of general relativity, mass bends light in such a way that one can observe objects that are further away than the object the telescope is pointed at. The *Compton Gamma Ray Observatory*, by contrast, operates on the assumption that the decay of different elements releases gamma rays at specific energies, which can be detected in a rather indirect fashion by making use of the Compton scattering that the gamma rays will trigger in the detectors.[48]

Neither one of these telescopes produces the one 'right picture' nor does any of them produce a wrong one. They rather all show different perspectives of the same object, argues Giere:

> Scientific observation is always mediated by the nature of instruments through which we interact with selected aspects of reality. In this sense, scientific observation is always perspectival. (Giere, 2006, p. 43)

Depending on the instruments used and how this data is evaluated, one will end up with a different 'picture' of the world, even with the same input.[49] That is, one will inevitably end up with different *models of data*. To compare the data to the theory, Giere (2006, pp. 68–69), following Suppes, argues that we only compare *models* of theory to *models* of data. That means, just as models of data represent a specific perspective of an object, so do models of theory represent a

[47] Furthermore, many telescopes do not even measure wavelengths within the visible range. And all the more for it because they give us information on the universe that we could not otherwise gain.

[48] The Compton Gamma Ray Observatory, which Giere gives as an example, was abandoned in 2000, but other observatories operate on similar principles.

[49] This, of course, is not just the case for telescopes but all kinds of scientific instruments. Another example that Giere discusses are brain scans. Depending on which technology (CAT, PET, MRI, etc.) is applied, a distinct image of the brain depicting different aspects of the brain is obtained (Giere, 2006, pp. 49–57).

specific point of view as well. A model is never a complete replica of the object or phenomenon it represents. It only has a *fit* that we consider good enough given the circumstances. Depending on the use of a model, the requirements the model has to fulfil will change. Giere (2006, pp. 72–81) compares this to the way we use maps. The same place can be displayed in very different ways depending on the map's purpose. Take, for example, the task of mapping the earth. Transferring a spherical object onto a two-dimensional map will naturally cause a problem for the geographer. While the well-known *Mercator projection* serves the purpose to give sailors a navigational tool, it also drastically distorts the actual relative size of different countries and continents depending on where they are located on the map. Greenland, for instance, appears to be more than 14 times bigger than its actual relative size, making it as large as the whole continent of Africa. Alternatives like the *Peters projection* or the *Robinson projection* correct for this particular problem but on the flip side have to make concession with respect to other aspects. The *Peters projection* shows the landmasses at its right proportions but with distorted shapes. The *Robinson projection* tries to combine the advantages of both those maps as best as possible but does so by curving the longitude lines and would, therefore, not be very useful for navigation. Thus, each of these different maps is useful for different objectives. They all show a unique perspective of the world: "representation is representation *for a purpose*" (Giere, 2006, p. 80). None of these maps can give a fully accurate representation of all aspects of the surface of the earth. Maps are not copies of the place they display. But then again that is, arguably, not the point of maps. Maps just like scientific models are, as Giere points out, tools that represent the world in a certain respect, in a way that is helpful to our specific (scientific) endeavour. This is by no means giving in to total relativism, as Giere emphasises. Scientists can very well determine that one of two of the same kind of instruments is faulty if they produce completely contradictory data.[50] Equally, when an object can be detected from different perspectives (e.g., with different instruments), then this can be understood, argues Giere, to mean that there

[50] To stay with Giere's example of the telescopes: if we were to point a gamma-ray telescope and a radio telescope at the same object and both deliver different observation, this would usually not be seen as a sign that one of the instruments is malfunctioning. But if instead there were two gamma-ray detectors that both are supposed to measure within the same range of wavelength registering something differently, scientists would, of course, conclude that at least one of the instruments is defective.

is “good evidence that there is *something* there, but this need not to be understood as knowledge in an ‘absolute objectivist’ sense” (Giere, 2006, pp. 57–58).[51] Meaning we cannot find the *one* ‘objectively’ true perspective of the object in question, in the sense of a ‘view from nowhere’.

3.2.3 Observations in climate science

The climate system is a global system. Collecting and processing climate data, thus, is a global task. Having access to global data is crucial in gaining knowledge and understanding of the climate system.

This is not a new insight. Predicting the weather has been an age-old human endeavour. Success in such diverse aspects of public life from farming to warfare are dependent on knowledge of how the weather will develop. In order to do so, since the 19th century scientists have tried to establish an infrastructure that would enable them to collect data on a global scale. It is, therefore, not surprising that meteorology – and subsequently climate science – were big data science early on and the first to develop “systems for producing globalists information” (Edwards, 2010, p. 24).

As noted in the introduction to this book, one line of argument frequently used by climate-change sceptics is that the models must be false because they seem to disagree with the observational data. Climate scientists, however, counter that observational data just as the models are affected by uncertainties.

Just as much as the climate system is complex, the observational data retrieved from it is also complex. Complex systems produce complex data in more than one way: for one, in respect to the amount of data and, for another, in respect to the methods of acquisition, processing and evaluation. Traditionally, observational data has often been viewed as providing a form of context-independent confirmation of theories. The example of climate science will show in the following how this separation clearly cannot be upheld in actual scientific practice. Observations are neither as independent nor self-

[51] Giere applies a very narrow and specific definition of *scientific objectivity* here. As we have seen in Chapter 2.3, this term has a rich history and even today a variety of interpretations. It might, therefore, be very well possible that Giere’s perspectivism is ‘objective’ in a different sense. An interpretation of *scientific objectivity* which does not imply a ‘view from nowhere’, similar to, e.g., Longino’s (1990) definition of objectivity as something that is achieved through diversity and a plurality of perspectives, would be much more compatible with Giere’s perspectivism.

vindicating as climate-change sceptics often claim. On the contrary, Paul Edwards shows that in climate science data is "model-filtered" and models are "data-laden" (1999, p. 437). The lines between theory or model, respectively, and observation are (at least to a certain degree) blurry. The complexity of the climate system lays bare the interdependency between those two sides of science traditionally treated as distinct. Creating global climate data sets requires much more than just 'collecting' data. Or to quote Edwards: "if you want global data, you have to make it" (2010, p. 321).

In the following, I will discuss what constitutes climate data, how this data is collected and processed as well as the difficulties arising in the process. I will then examine what this means for the evaluation of climate models with observational data and our understanding of the relationship between models and observations in general.

3.2.3.1 Climate data

What actually constitutes climate data and could we not just collect all weather observations from the last few centuries and be done with it? After all, weather observations have been made for centuries now.

Unfortunately, it is not that simple. One way to approach the difficulties of creating climate data is to look at the specific historical differences in the requirements for data in weather forecasting and climate science. As Edwards (2010, p. 292) points out what traditionally distinguishes one from the other is their purpose – in a nutshell it is a matter of speed versus stability.

The purpose of *weather data* is to forecast the weather of the next few days. Reliable, easily accessible data which 'arrives' within the time limits of the forecasting cycle is required. Weather data, which is retrieved only after the new forecast has been made, is of little use for making forecasts. By contrast, the purpose of *climate data* is to create a statistically useful account of the climate over a longer period of time. For this one does not so much require data that is accessible within a specific timeframe but shows consistency in the way it is collected over a long period of time.[52]

To visualise this difference, one might look at what meteorologist and climate scientists usually focus on when they discuss temperature. While meteorologists making weather forecasts are in search of the absolute temperature value of a specific moment in time, e.g., the temperature in Bochum tomorrow,

[52] For a good overview of the specific (historic) differences between weather and climate data, see Edwards (2010, pp. 294–295).

climate scientists, on the contrary, are traditionally interested in temperature anomalies, i.e., the deviation from the average temperature of a reference period.

3.2.3.1.1 Observations and uncertainties

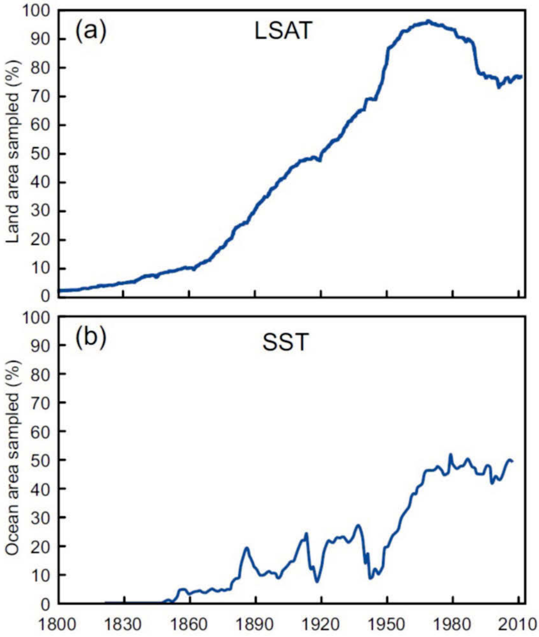
To establish a comprehensive picture of the change in the earth's climate, a huge amount data drawn from a variety of sources is necessary. Besides data from hundreds of land-based observational stations, climate scientists also rely, for instance, on ships and buoys, to obtain observational data from the ocean and radiosondes deployed on weather balloons, airplanes and satellites to get information about the climate from different altitudes (Chen et al., 2021, pp. 174–177). For insight into the state of the climate before the beginning of systematic observations in the 19th century, scientists also make use of proxy data, such as tree rings or ice-cores (Chen et al., 2021, pp. 177–178). The processes of creating homogenous data sets out of the different types of data are intricate and epistemically challenging undertakings.

In the following, some of the difficulties that arise from the complexity and plurality of the observational climate data will be highlighted. The objective is to show why the ideal of raw observations providing clear-cut, context-independent and 'objective' evidence of the validity of a model or theory is so detrimental, specifically in the context of climate science. As we will see, observations in climate science are distinctly theory-laden so that it can be really misleading to even speak of such a thing as 'raw data'.

Coherent, long-term observations of the past climate are paramount to understanding how the climate might change due to increased anthropogenic forcing.

Although first attempts of infrastructural coordination of weather and climate observation date back to the middle of the 19th century, up until fairly recently, the need of weather data has often taken precedence over climate data, as Edwards points out (2010, p. 287). The value of consistent long-term climate data was only fully appreciated in the second half of the 20th century. Scientists, for the most part, just did not foresee that there would be the need for dependable long-term data in the future and observations records were often not kept. Traditionally, meteorologists seldom had use for 'old' data in the process of weather forecasting and for a long time storage space for that amount of data had been very expensive.

Figure 4: Change in percentage of possible sampled area for land records (top panel) and marine records (lower panel). Land data come from GHCNv3.2.0 and marine data from the ICOADS in situ record



Source: Hartmann et al., 2013 p. 2SM-14 , Figure 2.SM.2

There are also often critical gaps in the recordings. Major historical world events can disrupt data records. There is, for instance, a lack of sea surface temperature (SST) data for the time of both world wars; see Figure 4. Most recently the COVID-19 pandemic has affected the amount of specific types of observational data collected during this time period because of a drop in air travel and ship traffic as well as interruptions in the regular maintenance of instruments. While the full effects of the pandemic on climate data records is not yet fully known, the consequences might not as bad as originally feared (Chen et al., 2021, p. 212).

Even when there is historical data at hand, there might also be gaps in information about the circumstances under which the data was collected, namely which specific instruments were used and whether there were environmental and structural circumstances that might have systematically influenced the data in one way or another.[53]

Fortunately, many inconsistencies are random and, therefore, cancel one another out over time. When, for instance, the temperature is measured at a significantly large number of stations, measurement errors will most likely deviate in both directions. But sometimes inconsistencies have systematic causes. One example is the overall change of position of instruments in the Alps during the 19th century from on and at buildings to open space (Böhm et al., 2001; Edwards, 2010, p. 299). Another case are systematic changes to the instruments used. A prominent example for this is the change in material used for the buckets employed to measure ocean temperature. The 'evolution' of buckets, from simple wooden ones, over canvas buckets[54] to modern insulated ones meant that there were temperature differences up to 1 °C, depending on seasonal and local variables (Folland and Parker, 1995).[55]

The specific circumstances under which observations were made can be difficult to reconstruct. In certain cases, scientists can refer to metadata, such

[53] For examples from sea surface temperature measurements, see Kennedy (2014).

[54] The wooden buckets are actually "relatively well insulated and tend to have larger volumes leading to smaller temperature changes" compared to buckets made out of canvas (Kent et al., 2010, p. 719).

[55] There are also other factors contributing to temperature differences in bucket measurements that have to be corrected for, such as "the size of the buckets (inner diameter and initial water depth for the canvas bucket model and bucket wall thickness for the wooden bucket model), the time the bucket was exposed on deck, the relative wind speed (which depends on the ship speed, the true wind speed and the degree of sheltering of the bucket) and the exposure of the bucket to solar radiation, all of which may vary from ship to ship and with time" (Kent et al., 2010, p. 723).

In some cases, measurements of SST are not taken by buckets but also through so called *engine room intake*, i.e., the seawater that is ultimately used to cool the engines of the ship. This kind of measurement also has a warm bias compared to buckets. Thompson et al. (2008) note that changes in the ships country of origin in 1945 led to an apparent temperature drop in SST: "Between January 1942 and August 1945, 80% of the observations are from ships of US origin and 5% are from ships of UK origin; between late 1945 and 1949 only 30% of the observations are of US origin and about 50% are of UK origin. [...] in August 1945 US ships relied mainly on engine room intake measurements whereas UK ships used primarily uninsulated bucket measurements" (Thompson et al., 2008, p. 648).

as user manuals of the instruments used and the likes, but those are not necessarily preserved (Edwards, 2010, pp. 317–319). Lack of this kind of information can constrict the accuracy of the 'picture' that can be drawn of the climate of the past from instrumental records.

Edwards (2010, pp. 17–28) argues that the creation of global-climate data sets is a question of infrastructure and insight into this infrastructure. Creating observational climate data sets requires having access to a globally organised network. From the 19th century onwards scientists slowly began creating a global observational network. Making use of this data requires climate scientists to do what Edwards calls an "infrastructural inversion" (2010, pp. 22–23): they have to turn the infrastructure 'upside down' to assess how the data was originally produced. However, even with an improved observational infrastructure, the problems described above are not purely issues of the past, as Edwards argues:

> Weather stations come and go. They move to new locations, or they move their instruments, or trees and buildings rise around them, or cities engulf their once rural environs. They get new instruments made by different manufacturers. Weather services change their observing hours and their ways of calculating monthly averages. These and dozens of other changes make today's data different not only from data collected 20 years ago, or even (sometimes) last week. It's like trying to make a movie out of still photographs shot by millions of different photographers using thousands of different cameras. Can we reconcile the differences, at least well enough to create a coherent image? Yes we can, scientists believe. But it isn't easy, and it is never finished. (Edwards, 2010, p. 6)

All in all, the need for long term, stable and global data make the creation of climate data sets a far form straightforward affair. Just collecting millions of single data points does not suffice; a great deal of data processing has to be done in terms of, reconstructing and homogenisation. Furthermore, climate modelling often requires gridded data meaning that the data points are spatially evenly distributed on a (virtual) grid. How this is done in practice calls for further methodological choices diversifying the approach that scientists can take (Parker, 2018).

3.2.3.1.2 Satellite data

Satellites as meteorological and climatological measurement facilities are a relatively recent invention. Images taken with the help of satellites were first used in the 1960s. But at first they were difficult to interpret and only marginally useful to local forecasting. Nevertheless, they were helpful in science communication as they provided visual aids for weather reports on TV, as Edwards (2010, p. 274) recounts. At the end of the 1960s, satellites carried the first instruments (radiometers) that were installed specifically to provide data for weather predictions. But it took a while until scientists had learnt how to implement this new data source into numerical weather predictions. Not until the 1990s, they actually substantially improved weather forecasting.[56]

These days satellite data are an essential part of weather forecasting and climate-change assessment. In the last decade, satellite data, notably a controversy about data from satellites equipped with so-called *microwave sounding units* (MSU), has also received some explicit attention from philosophy of science (see especially Edwards, 2010, pp. 273–279, 413–418; Lloyd, 2012). Even within the intricate sphere of climate data analysis, satellite data can be particularly complicated to 'read', which made integrating this data resource into meteorology and climate science such a difficult undertaking in the first place.

First of all, many instruments mounted on satellites only provide indirect measurements. Instead of temperature, MSU measure the microwave radiation emitted by oxygen molecules (radiance), from which then under the premise of a variety of physical and mathematical background assumptions the temperature of different layers of the atmosphere can be inferred.[57] Further, adjustments have to be made not just to filter out noise from the stratosphere but also to account for methodological issues. Especially considering that the MSUs are sequentially calibrated, the effects of orbital and instrumental decay have to be factored in (Wentz and Schabel, 2000). All of this comes with a variety of uncertainties and in practice requires complicated algorithms to account for those. That is, the 'raw' data is open to some degree

[56] For a more detailed account of the history of satellite data in weather forecasting, see Edwards (2010, pp. 274–276).

[57] To be more specific, the different frequencies of radiation are measured in distinctive 'channels' that can then be related, in a non-trivial way, to the temperature of different 'layers' of the atmosphere. In the case discussed here, scientists were interested in 'Channel 2' that measures radiance of the troposphere with some noise from the stratosphere.

of interpretations so that depending on the decisions made in this process different data sets can be obtained.

In the 1990s, an argument arose among climate scientists about the warming in the tropical troposphere. Models predicted that the tropical troposphere would show significant warming due to the increasing greenhouse gas emissions of the 20th century. But scientists Roy Spencer and John Christy (1990) claimed that the satellite data set developed by their research group, known as the UAH (*University of Alabama at Huntsville*) data set, evaluated with the help of radiosonde data actually disproved this. They argued that the radiosondes (on weather balloons) provided particularly reliable data as they are, contrary to satellite data, actually equipped with thermometers and measure the temperature of troposphere directly. However, other climate scientists, instead of discarding the model, questioned the reliability of the radiosonde data for the purpose of validating satellite data (e.g., Gaffen et al., 2000; Santer et al., 1999). They note that radiosondes are, in fact, prone to inconsistencies because they are exchanged frequently and their distribution is patchy. Lloyd (2012) points out that, contrary to what Christy and Spencer seemed to imply, radiosonde data does not provide a direct representation of 'reality'. The apparent evidence of cooler temperatures in the tropical troposphere that the radiosonde and UAH data sets were showing were misleading. Reconstructing MSU data in accordance with the radiosonde data set does not supply 'independent' evidence for the correct interpretation of the satellite data but relies instead on the (false) background assumption that radiosonde data could provide such.

In fact, it turned out that the same satellite data taking into account all methodological uncertainties could be processed in a way that created data sets that actually were much more compatible with the models: as was done in the case of the RSS (*Remote Sensing Systems*) and UMd (*University of Maryland*) data sets (Mears et al., 2003; Vinnikov and Grody, 2003). Eventually the dispute was settled, as far as the wider scientific community was concerned, at least to the extent that considering all uncertainties (observations and models) "there is no reasonable evidence of a fundamental disagreement between tropospheric temperature trends from models and observations" (Thorne et al., 2011, p. 66). However, research into these uncertainties continues to reduce the underlying shortcomings of both models and data sets (Chen et al., 2021, p. 175; Santer et al., 2017).[58]

3.2.3.1.3 Paleoclimate data and proxies

For information about the climate before the beginning of structured instrumental recording, scientists make use of so-called proxy data. This term refers to "any biophysical property of materials formed during the past that is interpreted to represent some combination of climate-related variations back in time" (IPCC, 2021a, p. 2245). From the width of tree rings, for instance, it is possible to infer whether it has been a particular warm or rainy year. Similarly, scientist can gather information about the state of the climate many millennia ago from the amount of oxygen, and the distribution of dust particles or pollen trapped in ice through ice-core drilling. Further scientists also make use of historical documents that go beyond weather station records.

Diaries, farmers' and ship logs, travellers' accounts, official documents and newspaper articles may provide information not just directly on the weather but also information on times of harvest, crop yield, droughts, frosts or vegetation in general which can give indications of past climate developments (Chen et al., 2021, pp. 177–178).

Further, recently, indigenous knowledge has been recognised more and more as a source of information, such as, e.g., Aboriginal knowledge about sea level rise in Australia passed on through oral traditions over 7000 years (Nunn and Reid, 2016).

So there are a number of sources for paleoclimate data even when there are no direct observational records, presenting scientists with information about climate variables some going back millennia. Still, particularly proxy data must be interpreted with care and reconstructing the climate of the past with the

[58] Santer et al. (2008), however, also concede that it might never be fully possible to solve the discrepancies in the observational data sets: "We may never completely reconcile the divergent observational estimates of temperature changes in the tropical troposphere. We lack the unimpeachable observational records necessary for this task. The large structural uncertainties in observations hamper our ability to determine how well models simulate the tropospheric temperature changes that actually occurred over the satellite era. A truly definitive answer to this question may be difficult to obtain" (p. 1719). Note, however, that this disagreement between models and observations is not considered by Santer et al. to mean that the models are necessarily wrong. The authors much more emphasise the uncertainties in observations.

help of proxy data hosts a variety of challenges. Here are just some of them (Frank et al., 2010; Parker, 2018; Schmidt, 2007):

For one, it can be difficult to accurately date proxy data. Some types of data has a yearly resolution (such as tree rings and ice cores), others can only be dated on a decadal scale (e.g., some pollen records or ocean sediment cores). Even with the more precisely datable data, there can be issues regarding allocating those data points to specific years, such as when it comes to the interpretation of tree ring growth (Mann, 2018).

Furthermore, most proxies are not equally well locally distributed. Ice-core data can only be sourced at the poles and tree ring growth is subject to seasonal differences.

There is also the issue that some proxies can be impacted by more than one factor. Plant growth for example, can be affected by temperature but also changes in soil and precipitation etc. So it is up to the scientists to figure out how to interpret the data and to find proxies that are less likely to be influenced by other factors.

And although proxy data is often calibrated against instrumental records of the recent past, sufficient instrumental records are only available from the last few centuries onwards, and the climatic circumstances of the earlier past might be outside the range of what we have instrumental records for.

While "sparse and noisy data are likely the underlying cause for the high methodological sensitivity" (Frank et al., 2010, p. 510) in paleoclimate records, proxies provide invaluable insight into past climate developments beyond instrumental records.

3.2.3.1.4 Reanalysis data

To fill the gaps of 'traditional' climate data, a new idea arose in the early 1980s to collect all available data of the last decades (or even centuries) and feed it into a weather model. The hope was to create a new, full-fledged data set that would provide a full "history of the weather, at every altitude, every grid point, every place on the planet" (Edwards, 2010, p. 323). After some years of searching for and assembling of data from all over the world, the first reanalysis projects started in the 1990s.[59] For this, climate scientists put this data retrospectively through a 4-D data assimilation as originally developed for weather

[59] For a more detailed historical account of the development of the first reanalysis projects from the idea to execution, see Edwards (2010, pp. 323–336).

forecasting.[60] The models used for this have to be frozen so that changes and improvements that are consistently done in weather modelling would not interfere with the process. The first data sets coming out of reanalysis projects covered only a timespan of between five and 35 years (Edwards, 2010, p. 326). These days reanalysis projects like version 3 of *20CR* (short for *Twentieth Century Reanalysis*) of the *National Oceanic and Atmospheric Administration* (NOAA) and the *Cooperative Institute for Research in Environmental Sciences* (CIRES) produce data for the years between 1836 to 2015 (Slivinski et al., 2019).[61]

Using data that has been produced by heavily relying on weather models to evaluate climate models has provoked initially some concern of philosophers that one would run into a problem of circularity "since weather-forecasting

[60] *Data assimilation models* originated in weather forecasting. The predecessor of these models, before numerical weather predictions (NWP) 'were a thing', was the so-called *analysis*. This consisted of handmade plotting of current data on maps from which scientists were literarily 'drawing', relying on their knowledge and expertise of the weather system, the forecast. With the introduction of NWPs gridded data became necessary. This gridded data was first produced by interpolating from the observational data by hand. Eventually, however, scientists began investing into so-called 'objective analysis', i.e., algorithmic process of interpolation. Then scientists started to integrate NWP forecasts as a 'first guess' for the time of observation. This had the advantage that scarcity of data in certain regions could be counterbalanced. Combining observations and forecast in *data assimilation* meant that uncertainties in model and data could be weighted and factored in. Soon scientists moved from three-dimensional assimilation to adding a fourth dimension: 'time'. This opened up the opportunity to integrate data lying outside of specific 'observing hours'.
Edwards emphasises that data assimilation has become much more than a "sophisticated version of interpolation" considering that "[a]ssimilation models are full-fledged atmospheric simulations; if run with no observational input at all, they would keep right on going day after day, month after month, generating physically consistent global data images. Where observations are available, they constrain the model, but they do not determine their output in any ordinary sense of 'determine'" (Edwards, 2010, p. 279). For a more detailed account of the development of data assimilation, see Edwards (2010, pp. 254–273).

[61] What timespan different reanalysis projects comprise depends, in practice, on the specific objectives, the available computing power and on what kind of data are used. The ERA-interim atmospheric reanalysis from the *European Centre for Medium-Range Weather Forecasts* (ECMWF) only goes back to 1979 when satellite data became available for assimilation (Dee et al., 2011). On the other hand, the 20CR data set, covering the years 1836 to 2015, is created assimilating only surface pressure observation in order to avoid issues of inconsistencies with the availability of observational sources for that timespan (Slivinski et al., 2019).

models include a number of assumptions about the physics of the atmosphere that are similar, if not identical, to those included in state-of-the-art climate models [...][and thus] the fit between reanalysis data sets and simulations of past climate [...] will be artificially inflated" (Parker, 2011, p. 587). Leuschner, however, argues that precisely the close relationship between climate and weather models provides a good argument not to be particularly concerned about reanalysis data insofar as "if an assumption works reliably in a weather model this can well be considered as an indicator for the adequacy of the assumption" (Leuschner, 2015, p. 370). She further notes that the assimilation models used for reanalysis projects are, for one, well-tested in their function in weather predictions and, for another, that reanalysis data are usually applied in conjunction with conventional climate observations.

Reanalysis data sets are now widely in use in climate science and are often treated and referred to the same as 'normal' observational data. From a philosophy-of-science perspective, Parker argues that reanalysis actually constitutes a form of complex measuring procedure, albeit it is a "measurement practice that is still under development" (2017, p. 294). She adds this caveat not because of a worry that there is something innately wrong with using computer simulations to produce data but because reanalysis data to date "are not subjected to a rigorous process of calibration that provides well-motivated uncertainty estimates" (Parker, 2017, p. 300) as one would expect for 'ordinary' measuring instruments and observation making. Parker particularly points out that part of the measuring process in reanalysis involves some difficulties that are specific to the use of computer simulations, such as numerical discretisation that is not part of conventional measuring practices and needs to be accounted for in order for reanalyses to be full-fletched measurement practices in their own right.

3.2.3.2 Model-data interdependency

What has been shown so far is that creating global climate data involves much more than simply collecting data from a variety of instruments at a variety of locations. These data sets are "models of data" (Suppes, 1962) and a great deal of work goes into constructing them. Before the 'raw data' is of any use to climate scientists, they have to be extensively processed. That is, climate data sets are, as Edwards calls it, *model-filtered* insofar as they are created with the help of "what we might call 'intermediate models' [...][which] include models of instrumental behaviour, interpolation techniques [...], techniques for automatic rejection of anomalous data points" and so forth (Edwards, 1999, p. 450). 'Raw' climate data on its own is patchy, inconsistent and sometimes conflicting. This is also a well-established insight in climate science: "[m]odel-filtered data can be trusted to the extent to which the models used to correct and extend the data have been independently tested and are confirmed" (Baumberger et al., 2017, p. 6). Or, to paraphrase Edwards: climate data has to be made (2010, p. 321).

However, Edwards also points out that the relationship between models and data is distinctly interdependent. Climate models are also "data-laden" (Edwards, 1999). Models are not just evaluated with the help of observations, they also contain in themselves a fair amount of observations. Specifically, the development of parametrisations requires scientists to consult observational data (Guillemot, 2010). For this reason parametrisations are also sometimes described as having a "semi-empirical" (Edwards, 1999, p. 449) character.

Tuning is another way in which models become data-laden. In the process of tuning models are calibrate with the help of observational data. There is some specific concern that this relationship might be questionable when the models are tuned to the same data they are later evaluated against. This practice, dubbed *double counting*, has sparked a discussion among climate scientists and philosophers. Scientists on the whole seem at least sceptical that this procedure could be considered adequate. Mauritsen et al. contend that evaluating quantities addressed in tuning are "of little value" (2012, p. 3) and the fifth IPCC Assessment Report (AR5) states that "quantities that are tuned cannot be used in model evaluation" (Flato et al., 2013, p. 749). A common strategy to avoid double counting in practice is *data splitting*: to use one half of the data set for tuning purposes and the other half for evaluation (Baumberger et al., 2017). Philosophers Katie Steele and Charlotte Werndl (2013), however, argue that from the point of view of Bayesian confirmation theory there is little difference between relying on the same data for tuning and evaluation and conventional methods of testing of hypotheses. Other philosophers and climate scientists (Frisch, 2015; Schmidt and Sherwood, 2015) have subsequently voiced criticism of this view. Frisch argues that, tuning has some confirmatory power but still concludes that "fit with data not used in tuning is a superior test of a model's performance" (2015, p. 174).

Edwards has called the connection between models and observation "symbiotic" (Edwards, 1999, p. 453). Contrary to the traditional ideal, there is neither a clear separation between data and model nor a clear-cut hierarchy where one of the two takes precedence. It is "a mutually beneficial but also mutually dependent relationship" (Edwards, 1999, p. 453):

> The picture that I hope is emerging here is that all knowledge about climate change depends fundamentally on modeling. It's not that there is no such thing as an observation separate from modeling. It's that putting together a trustworthy and detailed data image of the global climate – getting enough observations, over a long time span – requires you to model the data, to make them global. It's not that climate simulations models are perfectly reliable, any more than weather forecast models always get it right. Instead, it's that simulations already include a lot of data in their parameters; they are precisely not pure theories, but theories already partially adjusted to the conditions we observe in the real world. That's model-data symbiosis. (Edwards, 2010, p. 352)

This symbiotic relationship can be understood in two ways, Parker (2020) points out. On the one hand, it can be a mere reference to the instance that in general a model is created with the help of a data set which was created independently from this type of model, which in turn is used to process another kind of data set. But there is no direct, two-way exchange between one specific data set and one specific model. On the other hand, occasionally there are, Parker notes, also cases where the relationship is more direct – to the extent that one data set is created with the help of a model, which is then evaluated with the help of this specific data set.[62]

[62] One possible example for such a case, according to Parker, would be a model used to create synthetic data, which is then used to evaluate homogenising algorithm for finding, non-climate-change related inconsistencies in observational data. A direct symbiotic relationship may arise when the model producing the synthetic data is later evaluated against an observational data set, which was created with the help of the homogenising algorithm, which in turn was tested with the synthetic data. This kind of symbiotic relationship in general, argues Parker, does run a particular high risk of turning circularly in a self-affirming way, as the model "has no direct role in producing the climate data set; it merely plays a supporting role in efforts to evaluate methods for removing artifacts when producing the data set" (Parker, 2020, p. 815). Parker comes to the conclusion that reanalysis is one case where one might have more reason to be concerned that this relationship is problematic, to the extent that as "weather-forecasting models (used in data assimilation) and climate models take a similar approach to representing physical processes in the atmosphere, it could be that reanalysis data sets and climate simulations have some shared errors" (Parker, 2020, p. 816). However,

3.2.3.3 Verification and validation

One issue that arises in the context of evaluation of climate model concerns the applicability of the concept of *validation* and *verification*. Traditionally, the process of *verification and validation* (commonly shortened to V&V) is drawn upon to establish confidence in computer simulations:

> *Verification* is said to be the process of determining whether the output of the simulation approximates the true solutions to the differential equations of the original model. *Validation*, on the other hand, is said to be the process of determining whether the chosen model is a good enough representation of the real-world system for the purpose of the simulation. (Winsberg, 2019)

Both are commonly treated as distinct in execution and in conceptual classification. While verification is considered to be a question of mathematics and the accuracy of the numerical solution, validation concerns physics and the question whether the underlying equations of the model are an adequate representation of the target system. Both pose two separate questions:

> First, are the solutions that the computer provides close enough to the actual (but unavailable) solutions to be useful? [...] Second, do the computational models that are the basis of the simulations represent the target system correctly? (Frigg and Reiss, 2009, p. 602)

Although this conception is popular with scientists (Winsberg, 2018, pp. 156–157), philosophers have raised concerns that applying these terms to scientific models is problematic because "it is impossible to demonstrate the truth of any proposition, except in a closed system" (Oreskes et al., 1994, p. 641): a requirement that only purely logical or mathematical models can meet.[63] There is some disagreement within the philosophy of science community about the extent to which this concern about the entanglement of verification

Parker notes that current research does not show that reanalysis data sets are 'closer' to models than conventional observational data.

[63] Oreskes et al. come to the conclusion that in this instance instead of verification or validation the best that might be accomplished is *confirmation*, that is, there is an increasingly good match between increasingly diverse observations and the model output: "The greater the number and diversity of confirming observations, the more probable it is that the conceptualization embodied in the model is not flawed. But confirming observations do not demonstrate the veracity of a model or hypothesis, they only support its probability" (Oreskes et al., 1994, p. 643).

and validation holds in actual scientific practice and for which kinds of computer models (Jebeile and Ardourel, 2019; Morrison, 2015). Although it has been noted that practices concerning either primarily verification or validation are part and parcel of climate modelling (Winsberg, 2018, p. 157; Lenhard, 2018). Philosophers concerned with the peculiarities of complex computer simulations of the type of ESM have pointed out some features of these models that should make us sceptical about V&V as an epistemological concept of two separable procedures fully grounding our trust in these simulations. Why is that?

Lenhard (2018) argues that the applicability of V&V is limited in simulations of the type of global climate models because it would require to separate model structure and parameter. However, climate models necessarily also include parametrisations schemes with adjustable parameters, so that the adequacy of a model cannot be assessed without already having determined the parameter value. That is, "without assignment of parameters neither the question about representational adequacy nor the question about behavioral fit can be addressed" (Lenhard, 2018, p. 842). For this reason, he concludes that "[i]t is not possible to first verify that a simulation model is 'right' before tackling the 'external' question whether it is the right model" (Lenhard, 2018, p. 842).

Winsberg (2018, pp. 156–160) comes to the same conclusion from a slightly different angle. He points out that these models rather have a "life cycle" than undergo a "linear development" (2018, p. 158). Tuning, the need for parametrisations and a fuzzy modularity, means that the process of model development is an "iterative process" (Guillemot, 2010, p. 249; Winsberg, 2018, p. 158), where the model is consistently tested and further developed. New elements are added to the model, for instance, in form of new parametrisations and/or the discretization scheme is modified. That is, there are constantly changes made to both the underlying model and the implementation in a trial-and-error fashion, not just based on basic well-accepted physical and mathematical principles and theories, but also "physical intuition, phenomenology, local empirical finding, lore accumulated from parallel modelling successes, etc." (Winsberg, 2018, p. 158). Winsberg particular points out that this process leads to the possibility of unknown compensating effects so that:

> [w]hen a climate model succeeds at passing whatever test we subject it to, it might be because the underlying model is ideal and the algorithm in question finds solutions to that underlying model. Or it might be because of a "balance of approximations." This is likely the case when a model is delib-

> erately tailored to counterbalance what are known to be limitations in the schemas used to transform the model into an algorithm. [...] And when success is achieved in virtue of this kind of back-and-forth, trial and error piecemeal adjustment, it is hard to even know what it means to say a model is separately verified and validated. (Winsberg, 2018, pp. 159–160)

But Winsberg (2018, p. 160) also argues that this amalgamation of improvements efforts, constricting the model from different perspectives and objectives, are exactly what can substantiate the confidence that the models are adequate for a specific purpose.

Further, the notion that verification and validation cannot be kept fully separate does not mean that procedures targeting one or the other do not have an important place in the practice of climate modelling. As Lenhard notes the holism underlying the problem with V&V in complex computer simulations "comes in degrees" (2018, p. 842).

3.2.4 Conclusion

Contrary to conventional ideals about how science should operate this chapter has shown that in climate science the relationship between models and data is rather complicated and interdependent. For one, observational or experimental data is not just consulted at the end of the 'scientific process' to confirm or refute a theory or a model. Instead observations are a vital and intricate part in the climate-model building process. Not only are parametrisations often significantly based on observational data, but models are also continually calibrated, evaluated and further developed with the help of observations. These models are not just simply a representation of theory. They do not fit into the traditional theory-focussed narrative of science, whereby observations only play a minor character in the scientific process by verifying theories in the end.

Some philosophers like Hacking worrying that the experimental part of science was neglected have argued for the independence of the experimenter's work from that of the theorist's. However, when science is dealing with systems that are as complex as the climate system, it becomes increasingly undeniable that this separation into an empirical and theoretical part of science can no longer be maintained. Climate models are not just pure theory. They are laden with data through parametrisation schemes, tuning and evaluation processes. Similarly, observational data are clearly theory-laden. The pure, 'raw' measurement, the direct instrument reading without any processing, usually says little about the quality of the model.[64] Theory and observation are neither fully independent nor does one take precedence over the other. As Edwards (1999) has rightly pointed out, models and data have a "symbiotic" relationship. Importantly, this should not be confused with the claim that models and data would be interchangeable or in some way 'the same': "Interdependence is not identity; data sets are still derived *primarily* from observation, and models *primarily* from theory" (1999, p. 454).

Nevertheless, when the assumption of 'independent observations' providing irrefutable evidence for a theory or model is widespread in the public understanding of science, it makes it easier for science sceptics or particular interest groups to undermine trust in scientific research. Edwards (2010) and Lloyd (2012) have shown that in the case of the controversy about the interpretation of the MSU satellite data discussed above, the debate was not just held within the scientific community but also eagerly picked up by climate science sceptics from politics and the wider society as an argument against the reliability of climate models. The claim that the UHA data set, because it was calibrated with the help of radiosonde data, could serve as a kind of independent representation of 'reality', functioning as a benchmark against which the quality of models could irrefutably be judged, was also brought forward in hearings in front of the House Representatives in the United States. And the apparent 'mismatch' between the models and observations was presented as a failure of the scientists to do "sound science" (Edwards, 2010, p. 414, see also Chapter 3.3.1).

Instead what the case of the MSU data has shown is that, taking into account that observations just as models come with some degree of uncertainties, one mismatch with some specific data set does not necessarily mean that models ought to be disregarded right away.

Considering the controversy about the UHA data sets Lloyd (2012) introduces the distinction between *direct* and *complex* empiricism. From the point of view of direct empiricist 'raw' data are "windows on the world, as reflections

[64] In fact, Edwards notes that the meaning of *data* itself has changed in climate science. In the early days of climate modelling, scientists used to separate between data gained from simulations and that coming directly from observations. But this turned out to be "linguistically awkward" (Edwards, 2010, p. 283), so it became common practice to refer also to model output just as *data*. While many philosophers have been critical about the term *raw data* in general (Harris, 2003; Leonelli, 2019), computer simulations and techniques like reanalysis have expanded the definition of 'measurement' (Parker, 2017).

of reality, without any art, theory, or construction interfering with that reflection. This claim of a direct connection to reality is very important to their views" (Lloyd, 2012, p. 392). Lloyd argues that instead a "complex empiricism", which allows for the idea that data is theory-laden, in need of extensive processing and is to a certain degree open to interpretation, would be more appropriate.

But, of course, despite all these difficulties, climate models are, nonetheless, broadly and continuously evaluated in respect to their fit with observations. But as observational data sets are, on the one hand, model-filtered and, on the other hand, models are data-laden in terms of tuning and semi-empirical parametrisations, it can be difficult to determine what a 'good' fit of models with observational data actually means. In this respect, climate scientists set much store by the *adequacy-for-purpose* principle (Parker, 2009; see also: Notz, 2015; Knutti, 2018; Chen et al., 2021, p. 221). While in general the models constantly get better at representing the climate of the past, they, nevertheless, do not, as we have seen in Chapter 3.1.3, display all relevant variables and processes equally well. So it has to be established what specific variables and processes are relevant for the model to be adequate for a specific purpose. Thus, "[t]he challenge [...] is to determine which instances of fit do support and which instances of misfit do undermine an adequacy-for-purpose hypothesis" (Baumberger et al., 2017, p. 6). However, even if the properties of fit a model is required to display for a specific purpose are determined, a model showing a good fit in this respect still does not necessarily warrant that the model will be adequate for making predictions about the future, as:

> [i]nstances of fit could be the result of compensating biases, or overfitting, or could simply be unimportant if the evaluated quantity is unrelated to the prediction of interest. Instances of misfit could result from the fact that the model simulates a different quantity than that observed or from biases in observations, or of different processes in models and observations. (Knutti, 2018, p. 331)

This means, that it is not possible to simply extrapolate from an adequate representation of current and past climate that the models will necessarily be able to represent future climate states with high anthropogenic forcing equally well. That is, the climate of the future might lie outside the boundaries of what we currently have data for, so that feedbacks[65] or other processes might emerge that are not accounted for by evaluating the model with regard to their fit to the available data (Baumberger et al., 2017, pp. 8–9).[66] Further, it might very well be that (unknown) compensating effects within the model contribute to a good fit to observations that might not hold under future climate change conditions because, for instance, feedback processes might accelerate in unforeseen ways.

On the other hand, when a model is not able to represent a feature of the climate of the recent past, then it can be said that it is highly unlikely that the model is adequate for projecting that feature of the climate in the future (Parker, 2009). Therefore, a good fit of the model with observational data has become a necessary but not sufficient condition for the model to be adequate for the particular purpose. In other words, "empirical accuracy of model results should [...] be understood as premises" (Baumberger et al., 2017, p. 7).

For these reasons, a good understanding of different climate processes and the system as a whole are considered by climate scientists key to estimating to what extent these processes are adequately represented in the models (Bony et al., 2013, p. 20). Based on this assumption, scientists specifically highlight the necessity of a proper understanding of the inner workings of the models themselves when one wants to draw any conclusions about the models' applicability for climate change predictions:

> We need to make sure the models do the right thing for the right reason, because we want to use them beyond the range they have been evaluated. We have greatest confidence in models where we understand the processes behind the results, and where we can argue that models represent them well enough. (Knutti, 2018, p. 346)

Knutti also argues that "process understanding" (2018, pp. 334–338) is central to the adequate-for-purpose question. That means having an insight into whether the emergent component of the model arising out of the inner model

[65] Climate feedbacks refers to those processes in the climate system where a change in quantity *a* also impacts another quantity *b*, which leads to further change in quantity *a*. In general these interactive relationships can, depending on the specific process, have an accelerating (positive feedback) or decelerating (negative feedback) in the context of climate change. Important feedbacks concern, for example, clouds, the carbon cycle and ice-albedo.

[66] One way that climate scientists deal with these problems is establishing so called *emergent constraints* (this will be further explored in Chapter 3.3.3.4).

structure, are similar to the target system. Further, process understanding requires that the interactions of the emergent elements in the model are well enough understood, so it is reasonable to believe that, firstly, it will remain like that over a significantly long period of time and outside the range that can be evaluated and, secondly, that no other relevant feature is missing. Only if all of this is accomplished, argues Knutti, one could consider a model adequate for purposes that lie outside the range for which the models can be tested. Thus, mitigating the epistemic opacity of the models through background information that climate scientists have acquired both about the inner-workings of the model and the target system and understanding to what extent specific climate processes are adequately represented in the model is seen as an essential aspect in improving climate model projections (see also Baumberger et al., 2017).

All of this requires much more than pure model-data comparison. Instead, as will be further examined in Chapter 3.3.3.4, what makes climate scientists confident in using models for specific purposes are usually a variety of different factors that go well beyond simply being in accordance with observations; the model fit is just one aspect among many.

To summarise it can be said that what has been stated in this chapter about the relationship between the empirical and theoretical part in science holds, specifically for instance, where science deals with highly complex systems. But it would be an illusion to say that these two spheres of science can be fully separated, even when science is concerned with less complex systems. And as science turns to more and more complex problems, this will just become more and more obvious. On the other hand, as we have seen Chapter 3.2.2, philosophers and sociologists have long been pointing out the complex role of observations in science and climate science is here no different than other sciences (Guillemot, 2010).[67]

One reason why the ideal that observations provide a readily available irrevocable benchmark against which theories can very easily be evaluated has to be assumed to be rooted in the fact that so far the need for and difficulties of processing and handling of data has not been particularly noticeable from

[67] Guillemot particularly notes that compared to many other sciences the "data malleability is of a much higher degree in the climate sciences, due to the extensive use of computers on all levels" (Guillemot, 2010, p. 249).

outside of science. The wish to clearly separate one from the other is understandable. Vindicating climate models would be much easier and debates about the specific impact as well as scale of anthropogenic climate change could be settled much quicker if there was some kind of irrevocable method of providing evidence in the form of observations telling us instantly if the models were right or wrong. If we come to the table with these expectations, then it seems altogether unsatisfying to hear that the only possible statement about the model is whether increased confidence in the adequacy of the model for a specific purpose is warranted. But things are rarely as neat in science as outsiders and often even scientists themselves wish them to be. Science is messy and complicated. This is not a denial of climate change nor a dismissal of science itself. As illustrated above in the case of the MSU data, even in situations where there is disagreement on the interpretation of the data at first, over time a consensus will be found within the community. Scientists, in the end, are often able to settle these kinds of debates. But this takes time as well as effort and will always be accompanied by some (never fully disappearing) uncertainties.

3.3 Predictability

3.3.1 Introduction: predictability and uncertainty

The wish to predict the future, to produce forecasts of what is to come is an age-old human endeavour. In ancient Rome, animals, specifically birds, were observed to divine some knowledge about the will of the gods. In ancient Greek, the oracle of Delphi was consulted, before wars were fought, and astrological readings of the stars can be found in many cultures going back millennia.

As far as science is concerned, making predictions as an integral, central function of science turned into an essential part of the definition of science in the 17th century, including "the conviction that any discipline that does not make successful predictions thereby fails to make good its claims to providing scientifically adequate understanding" (Rescher, 1988, p. 25). The emphasis on prediction making in this conception of science only gained in intensity over the next centuries with Newtonian physics and enlightenment being driving forces. The hope was that science would progress to supply ever more precise predictions. However as Rescher notes, the rise of the predictability ideal in science, took a hit in the wake of the French Revolution and only came back "as science of nature gave way as source of inspiration to the science of society" (Rescher, 1988, p. 25) in the 19th century.

In the post–World War II period, the public and political 'appetite' for predictions reached a new high (Rescher, 1988, p. 28). At the same time advancements in computer power and the rise of computer simulations in the second half of the 20th century increased the abilities of science to fulfil these expectations. The first weather models came out of the effort put into simulating the explosion of nuclear weapons at the beginning of the Cold War when John von Neumann recognised that the insight acquired there could also be applied to weather forecasting (Weart, 2010, p. 209).[68] The use of computer simulation spread to multiple other fields of science over the rest of the century. Computer simulations now have a wide variety of applications, such as economics, epidemiology, engineering, cosmology and much more. Further, the progress in computer modelling did not only meet the request for information about the future, it also facilitated it, as Heymann et al. point out: "models helped both to create and furnish social demands for predictive knowledge" (2017b, p. 6).

However, while at once the demand for predictions increased from the middle of the last century onwards, at the same time, with this increasing relevance of science in public life, the pressure for science to provide clear, irrefutable or, to be more specific, uncertainty-free predictions also increased. From the middle of the last century onwards it became a popular and very effective strategy of some interest groups to undermine research results inconvenient to them, by arguing that there are still too many uncertainties to take actions, and that, before any action can be taken, there first needs to be more research done (Oreskes and Conway, 2010).

A helpful way of thinking about these two somewhat contradictory expectations about what science ought to deliver is in terms of the concept of *cultures of predictions* as applied by Heymann et al. (2017a).[69] This framing will be use-

[68] It still took a long time from that initial idea to any kind of weather simulation that could be used for actual forecasting purposes. Two names that should be mentioned here are Jules Charney, who was engaged by Neumann to lead the group that would develop the first weather model and Lewis Fry Richardson, who developed the numerical system based on Bjerknes equations (Chapter 2.1), which was used to build the first weather model (Richardson, 1922). For a longer recount of the history of weather and climate modelling see Weart (2010) and Edwards (2010).

[69] The concept of "culture[s] of prediction" was not invented by Heymann et al. (2017a) as the authors themselves state. For instance, Fine (2010) first applies the phrase "cul-

ful to understand the way the social and the scientific spheres are interconnected through reliance on the predictiveness of science, the difficulties that come with the interdependency of these two spheres and how this specifically translates to the case of public perception and expectation of climate science. Heymann et al. characterise the cultures of prediction in five steps in order to show "the broad-ranging and pervasive role of predictive efforts in postwar modern society" (2017b, p. 6):

1. the social role of prediction;
2. the character and significance of computational practices;
3. the domestication of uncertainty;
4. the degree of institutionalization and professionalization of predictive expertise;
5. the cultural impact of predictive practices and claims (Heymann et al., 2017b, pp. 6–7).

Firstly, predictions fulfil a significant social role. The general expectations for science to provide knowledge about what the future will bring extend to most aspects of society. Heymann et al. (2017a, pp. 20–22) note that this also means that prediction making can sometimes take precedence over understanding the system, a prioritisation that is not uncontroversial within the scientific community. This is also a discussion that has been taking place in climate science (Bony et al., 2013). However, it has to be said that, as will be further discussed below, it is questionable whether understanding the systems and the simulations, and improvements in predictions making skills can be fully separated. Consequently, some climate scientists advocate to put the focus of climate science towards finding ways to mitigate the epistemic opacity of the models (e.g., Baumberger et al., 2017).

Secondly, with the vast demand for predictions, scientists increasingly rely on computer simulations (Heymann et al., 2017a, pp. 22–26). This comes along with specific epistemic challenges, many of which have already been discussed in the previous chapters (see particularly Chapter 2.1).

ture of prediction" in his in-depth study of American Weather Forecasting Institutions, while Johnson (2017) uses the term to explore the role of mathematics in prediction making. Heymann et al. say they specifically use the plural here to emphasise "the local origin and socially contingent character of the cultural formations built around the construction and use of computer models for predictive purposes" and that there is "a multitude of distinct cultures of predictions" (Heymann et al., 2017b, p. 6).

Thirdly, prediction making usually comes hand in hand with uncertainties science has to deal with alongside the public's cravings for certainty (Heymann et al., 2017a, pp. 26–29). The specific kinds of uncertainties climate modellers face will be discussed further in the following. As already noted in the introduction to this book, in the context of climate science uncertainties often get misrepresented in public debates as 'evidence' that nothing can be said about anthropogenic climate change at all. The question of how to deal with uncertainties also has to do with the first characteristic of *cultures of predictions* as climate scientists argue that uncertainties can be reduced by increasing background knowledge about the models and the target system (Bony et al., 2013; Knutti, 2018).

Fourthly, roughly in the last seventy years, the increasing demand for scientific predictions has also led to the creation of a variety of institutions whose primary concern is to provide predictive knowledge. In the case of climate science, arguably the most significant one of those institutions is the IPCC, whose role in assessing and communicating climate change will also be discussed below (Heymann et al., 2017a, pp. 29–32).

Last but not least, the focus on predictions in the last century also had a wider cultural impact. One prominent example is how much the public mindset has been directed towards climate change. The cultural impact and the power are such, argue Heymann et al. (2017a, pp. 32–36), that we are usually not aware how the cultures of prediction permeate everyday life and direct our view of the world.

Heymann et al. argue that "cultures of prediction represent cultures of power" (2017b, p. 7). However, they also note that the power attributed to science in its abilities to make predictions, does not necessarily translate to political actions. On the contrary, it can halter or even undermine political actions. In the context of climate science, this can be exemplified with help of the phrase 'sound science'. It relates to an argument against taking up a stronger climate-change mitigation policy, brought forward by climate science sceptics. That is that there are still too many uncertainties and that, unless they are eliminated, it would be an overreaction to act. As Oreskes and Conway (2010, pp. 136–163) have shown, the method to call for 'sound science' as a premise before any political action should be taken first arose as a strategy of tobacco companies in order to discredit research showing connections between second-hand smoking and cancer, and to avoid further regulations in the early 1990s. A strategy that was quickly adopted and reapplied to argue against, among other issues, regulations of CO_2 emissions, as Oreskes and Conway further show (pp. 169–215). By the mid 1990s, the claim that climate scientists were not doing "sound science" but "junk science" found its way to hearings at the US House of Representatives and Senate (Edwards, 2010, pp. 411–414). In Chapter 3.2. we have seen that the demand for 'sound science', alluding that there is an apparent discrepancy between observations and models, can function as an embodiment of the ideals that observations provide irrefutable evidence for or against the models. On the other hand, the call for 'sound science' also puts the emphasis on the uncertainties that come along with making predictions. As Supran & Oreskes (2017) have shown in an analysis of internal communication of Exxon Mobile Cooperation insinuating that there is not yet sufficient evidence of global warming was a well-established tactic to thwart stricter climate policy. They conclude that in order to avoid regulations the company relied on a publicity strategy that "overwhelmingly emphasized only the uncertainties, promoting a narrative inconsistent with the views of most climate scientists, including ExxonMobil's own" in order to "undermin[e] public understanding of scientific knowledge" (Supran and Oreskes, 2017, p. 15).

Thus, while the post–World War II period increased the relevance of science in the public sphere and the expectation but also ability (through rising computing power) to provide predictive knowledge, it also brought with it the expectation of science to deliver binary, clear-cut and uncertainty-free predictions. Thereby, science risks becoming trapped in the conflict between the self-awareness that knowledge is preliminary and the public expectation to provide clear-cut answers to questions of the future. When the misconception that uncertainties are a sign of bad science or not fully matured science is widespread in the public perception of science, this can quickly become a problem. While scientists whose work is under public scrutiny, like climate science, are often very much aware that the complexity and the impact of their research does warrant caution, as the following will show, they also have to navigate, on the one hand, public expectations to provide meaningful answers and, on the other hand, the knowledge that any misplaced overconfidence on their part can be misconstrued as a sign of the overall corruption of their field of research.

Considering the structure of this subchapter, I will first discuss *robustness analysis* (RA) as it has become central to the debate in philosophy of science on how to deal with the uncertainties occurring in modelling and specifically in climate modelling. I will, then, go on to discuss uncertainties in climate science more explicitly, how they are communicated to policymakers and stakeholders and how the argument about RA plays out here. This analysis of uncertainties and RA, on the one hand, will show that the sources of uncertainties climate scientists are confronted with are multifaceted and intricate so that the demand for climate science to reduce these uncertainties is often easier said than done. But, on the other hand, this does not mean that these uncertainties fully debilitates climate science and render it impossible to make any kind of statement about the way the climate will alter due to anthropogenic forcing. It will be shown how, despite all kinds of uncertainties, it is possible for climate scientists to assess specific hypotheses. This will be done by way of example of Equilibrium Climate Sensitivity (ECS) which is one of the most important variables to determine the effects of anthropogenic climate change.

3.3.2 Robustness

The notion of RA of models was first introduced in philosophy of science by Richard Levins (1966). As a biologist, Levins observes that many models in population biology, his field of research, include idealisations and simplifications due to the complex nature of the systems in question. This, in turn, means that the models also include elements that are not 'truthful' representations of the world.

Because the systems are often too complex to model these systems with a "naïve, brute force approach […] which is a faithful, one-to-one reflection of this complexity" (Levins, 1966, p. 421) to be feasible in practice, the question arises how we can still infer knowledge about a specific phenomenon from the models despite these idealisations. That is, how do we know that a model result is due to "the essential of a model or […] the details of the simplifying assumptions" (Levins, 1966, p. 423).

Levins' proposal for a solution to this conundrum is to use a variety of different models with different idealisations that, nevertheless, share a common core regarding the phenomena:

> we attempt to treat the same problem with several alternative models each with different simplifications but with a common […] assumption. Then, if these models, despite their different assumptions, lead to similar results we have what we can call a robust theorem which is relatively free of the details of the model. Hence our truth is the intersection of independent lies. (Levins, 1966, p. 423)

Levins concludes that, when models complement each other to the extent that they are "coordinate alternative models for the same set of phenomena" (1966, p. 431) and generate matching results, this constitutes a *robust theorem*. Meaning a common, reliable prediction, even though every single model is wrong in respect to the representation of some aspect of the target system.

Levins' account of robustness has subsequently been criticised by Orzack and Sober for asserting "that a statement's *robustness*, as distinct from its *observational confirmation*, can be evidence for its truth" (1993, p. 538). They criticise Levins' concept of RA for seeming to provide, as Weisberg put it "a novel, nonempirical form of confirmation" (2006, p. 732). Orzack and Sober provide a more formalised account of RA. They begin by identifying the specific circumstances under which, they say, a clear relationship between robustness and truth can be established:

> There is a special case in which the connection between robustness and truth is clear. Suppose we know that one of a set of Models M_1, M_2, ..., M_n is true, but we do not know which. If R is a robust theorem with respect to this set, then R must be true. (Orzack and Sober, 1993, p. 538)

However, while in this specific situation the premise is that one of the models in the set is true, Levins' assumption is that all models are "lies" one way or another. Orzack and Sober argue, that unless we know that at least one of the models in the set is true, we cannot infer from the fact that the models predict the same that it is true.

Otherwise it might very well be the case that a robust theorem arises because of a common denominator in the form of a shared assumption within the model, which might or might not be a lie that all models have in common. From this point of view, a robust theorem might reflect more about the conveniences of the model building process than its truth (Orzack and Sober, 1993, p. 538).[70]

[70] Orzack and Sober also consider what independence of models means. Like Levins they regard the independence of models in a set as a necessary premise for robustness (1993, pp. 539–540). They argue that there are two ways in which the models could be considered independent: logical or statistical. With regards to the first case, it has to be said that "competing models are not logically independent" (Orzack and Sober, 1993, p. 539), and with regard to the second case, there remains the question of how one would sample from the whole set of models. That is, reason Orzack and Sober, both kinds of independence cannot be applied to RA of models in scientific practice.

Levins, subsequently, has defended his approach to robustness from Orzack and Sober's attack by pointing out that RA as he understands it does not forego empirical observations:

> Orzack and Sober are worried that the robustness strategy seems to propose a way to truth independent of observation. This is not the case. Observation enters first in the choice of the core model and the selection of plausible variable parts, and later in the testing of the predictions that follow from the core model. (Levins, 1993, p. 554)

The way RA is represented by Orzack and Sober does not take this into account, Levins argues.[71] He notes that seeking robust theorems "reflects the strategy of determining how much we can get away with not knowing, and still understand the system" (1993, p. 554), though as Weisberg points out Levins "does not tell us how it helps to confirm models and their predictions" (2006, p. 732).

Building on this discussion, Michael Weisberg (2006) offers a new approach to RA. Weisberg's goal is to show that "robustness analysis is effective at identifying robust theorems, and while it is not itself a confirmation procedure, robust theorems are likely to be true" (2006, p. 732). To do so requires a more differentiated understanding of the concepts of *robustness theorem* and *robustness analysis*, argues Weisberg.[72] To that end he proposes to see RA as the following four-step procedure (Weisberg, 2006, pp. 737–738):

1. Finding a robust property that is a shared result among an ensemble of models
2. Studying the models to find a common structure that creates the robust property

[71] Lloyd, whose account of RA in the context of climate models will be discussed further below in Chapter 3.3.3.4, argues that the differences between Orzack and Sober, and Levin are the divergent objectives: "Orzack and Sober had a different goal, namely predictive inference to the model's outcome […] about which they were likely correct […]. Levin, in contrast, emphasised the key empirical evidence for the model structure under consideration" (Lloyd, 2015, p. 59).

[72] Weisberg presents his concept of RA in the context of the example of a predator-prey model. I will just introduce his argument here in an abstract form as I will discuss an application of Weisberg's four-step robustness scheme to climate modelling by Lloyd (2015, 2010, 2009) in Chapter 3.3.3.4.

3. Determining how the given mathematical model structure is to be empirically interpreted
4. Undertaking stability analyses by examining how the robustness theorem will fare if the models in the ensemble change somewhat

The first two steps, argues Weisberg, go hand in hand. Once (or often while) a robust property (that is a common predictive result the models generate) is found in "a sufficiently diverse set of models" (Weisberg, 2006, p. 737), the core structure which brings about the robust property has be determined. Weisberg calls this the *common structure*. In the most basic case, the common structure has the same mathematical form in every model, but this is not necessarily the case.[73] After the first two steps a specific mathematical description has been obtained but not yet any connection to an empirically observable phenomenon has been made. This follows in the third step. Without determining this empirical description, the robust property might as well be just a mathematical construct to be found in all models but does not tell us anything about the real-world system we are interested in, Weisberg points out.

These three steps culminate, he argues, in the formulation of the *robustness theorem*, which has the following general conditional form:

> Ceteris paribus, if [common causal structure] obtains, then [robust property] will obtain. (Weisberg, 2006, p. 738)

In a concluding fourth step of Weisberg's definition of RA, different kinds of stability analyses are performed in order to investigate what happens to the robust theorem when the circumstances characterized in the models change slightly.

This is, according to Weisberg, the four-step process of RA. But Weisberg also states that RA has some epistemic power. However, he emphasises that, while RA is applied in science to further knowledge and understanding about real-world phenomena, determining a robust theorem by itself is not sufficient:

[73] Weisberg adds that, in those cases where the common structure has not the same mathematical form, it can make the identification and analysis of the common structure much more difficult and may lead to situations "in which theorists rely on judgment and experience, not mathematics or simulation, to make such determinations" (2006, p. 738).

> A common reason theorists engage in robustness analysis is to increase the quality of their predictions and explanations about real phenomena. Although useful for both of these purposes, the theorems generated by robustness analysis cannot fulfill either role alone because they are conditional statements, further attenuated with ceteris paribus clauses. Explaining a real-world phenomenon or predicting its occurrence requires us to know that the common structure is actually being instantiated and that no other causal factor is *preempting* the efficacy of the common structure. (Weisberg, 2006, p. 739)

To do so, one would usually turn to empirical testing. However, Weisberg notes that RA is often done in instances where options to do so are lacking. He sees RA as a procedure that under certain circumstances is still informative and "can give us good reasons to believe the predictions and explanations of robust theorems" (Weisberg, 2006, p. 739). To be, thus, epistemically informative, two questions have to be answered, Weisberg argues:

> 1. How frequently is the common structure instantiated in the relevant kind of system?
> 2. How equal do things have to be in order for the core structure to give rise to the robust property? (Weisberg, 2006, p. 739)

In the absence of empirical data, the first question can be answered, at least to a certain degree, by making sure that a "sufficiently heterogeneous set of situations is covered in the set of models" (Weisberg, 2006, p. 739). Weisberg argues that, once it has been determined that a satisfactorily large and varied number of models show the same causal structure and respective robust property, then one can assume that it is likely that the same causal structure is at play when the aforesaid robust property is detected in the real world.

The second question is dealt with in step four of RA, where the question to what extent the robust property is stable under varying background assumptions is addressed.

Weisberg's goal is to show that Orzack and Sober's concern that RA promises a non-empirical kind of confirmation is unfounded. He notes that the third step in his interpretation of the RA-process constitutes the kind of jump from a pure mathematical statement to an empirical one, which Orzack and Sober criticised. However, Weisberg counters that this process "is actually part of a well-accepted theoretical practice that is so common, it is rarely discussed explicitly" (Weisberg, 2006, p. 740). He argues that there is often an implicit step in the confirmation process before the empirical confirmation, where the scientists ask the question whether, if the causal connection actually holds in the real world, the model would actually be able to represent this process appropriately. Weisberg calls this "low-level confirmation":

> Despite rarely being discussed explicitly, theorists' confidence in their ability to represent phenomena with their models did not come for free. It was minimally established by demonstrating that the relevant mathematics could be deployed to make correct predictions. It may also have been investigated explicitly by mathematicians. These investigations result in what I will call *low-level confirmation,* confirmation of the fact that certain mathematical structures can adequately represent properties of target phenomena. (Weisberg, 2006, p. 740)

Weisberg argues that in his conception of RA low-level confirmation is an element of the third step. It allows us, he argues, to draw a conclusion about the causal relationship between the robust property and the model structure. That is, low-level confirmation is helpful to make the step from a description of a pure mathematical relationship to some empirical assumption.

Herein lies for Weisberg the strength of RA. In those cases where empirical confirmations are difficult due to the complexity of the system, "it identifies hypotheses whose confirmation derives from the low-level confirmation of the mathematical framework in which they are embedded" (Weisberg, 2006, p. 741).

Although Weisberg does not give it a great deal of attention, much of his argument hinges on the condition of a "sufficiently heterogeneous set of models" (2006, p. 739). However, the subsequent discussion (Schupbach, 2018), specifically about the applicability of RA to climate modelling (Lloyd, 2010; Parker, 2011; Winsberg, 2018), centre on the question of what constitutes "sufficiently heterogeneous".

Before I will turn to this debate, I will sketch out what kinds of uncertainties climate scientists face and how climate science deals with this, to see why some philosophers of science have voiced doubt whether RA as outlined above can be applied to climate modelling and if there are alternative routes to establish robust evidence for climate science hypotheses.

3.3.3 Uncertainties in climate science

There are many sources of uncertainties in climate science. This is not to say that there is actually serious doubt in the scientific community that anthropogenic climate change is indeed happening or that its consequences, even with conservative estimates of temperature rise, would not be severe, as every new edition of the IPCC report demonstrates. After all, many relevant processes of the climate and climate change are well understood. Particularly, the link between an increase of carbon dioxide emission and the rise of General Mean Surface Temperature (GMST) has been known about for well over a century (Arrhenius, 1896). Nevertheless, uncertainties play a major role in the public climate-change discourse. So before turning to what makes climate scientists confident in their work despite uncertainties, let us take a look at what causes these uncertainties and how they are communicated.

Broadly speaking sources of uncertainties in climate modelling can be categorised into three different types (Lehner et al., 2020):

1. Model uncertainty
2. Climate variability uncertainty
3. Scenario uncertainty

Model uncertainty is of an epistemic nature (Knutti, 2018, p. 329); that is, not inherent to the system but arises due to our lack of understanding of and lack of means to represent the climate system. These will be explored in more detail below divided into structural and numerical uncertainties, and parameter uncertainty.

On the other hand, uncertainties coming from the internal variability are innate to the climate system. The climate system does not just deviate from the mean state due to external anthropogenic (e.g., greenhouse gases) or natural forcing (e.g., volcanic eruptions) but also internal (e.g., the El Niño-Southern Oscillation) processes. When it comes to assessing climate change with the help of models, internal variability becomes an issue, particular on a shorter time scale, when (disregarding all model uncertainty) it is not clear if a change in the climate is due to external forcing or some random internal variability. These effects can be dealt with by running the model multiple times while varying the initial conditions[74] and by averaging the relevant climate variables over a longer time period. However, due to the computational costs of making enough model runs to sufficiently explore internal variability it can remain a significant source of uncertainty particularly for shorter timescales.

I will not discuss uncertainties about possible emission scenarios here in any detail. However, they are by no means unimportant. Quite the contrary, as climate models only project how the climate will change under specific emissions scenarios, they are vital to questions of climate-change policy. But these are uncertainties that are directly dependent on human behaviour, not natural laws and lack of knowledge thereof and, therefore, will not be examined here in detail.[75]

Climate models are also not free from the effects of observational uncertainties considering the "symbiotic" nature of the relationship between models and data (Edwards, 2010, 1999). The variety of sources of data uncertainties has been examined in Chapter 3.2.3 and should also be kept in mind in the context of issues concerning model evaluation.

[74] When it comes to climate projection as compared to predictions, it is commonly said that these are independent of initial conditions as models for projections are not run from observation-based initial conditions but from an assumed preindustrial state. As climate projections are used to explore the impact of external forcing on the climate system in general, the specific initial conditions are less relevant. Internal variability is sometimes nevertheless referred to as *initial-condition uncertainty* in the context of projections – when there is uncertainty regarding the question if the model spread is due to external forcing or 'normal' internal variability because the model performance has not been explored systematically enough in respect to varying initial conditions. However, from a philosophical perspective to what extent it can be said that climate projections are or are not affected by initial conditions uncertainty hinges on the precise definition of the term as Werndl points out. (For internal conditions in the context of predictions and projections, see Werndl, 2019; for the difference of the significance of internal variability to weather forecasting and climate projections, see also Winsberg, 2018).

[75] Climate scientists usually consider human behaviour to be external to the climate system. One might speculate whether this has a more fundamental reason that goes beyond pure practical considerations about the models as Parker writes: "Classifying human activities as external to the climate system seems to be a pragmatic choice—it is easier, and a reasonable first approximation, to represent anthropogenic greenhouse gas emissions as exogenous variables in climate models—though it may also reflect a deeper ambivalence about whether humans are part of nature" (Parker, 2018).

3.3.3.1 Numerical approximation and structural uncertainty

No climate model is a perfect and complete copy of the real climate system. There are two reasons for this.[76] On the one hand, climate models are numerical models. At the core of ESMs and similar models are fundamental partial differential equations (see Chapter 2.1). The problem is that these equations cannot be solved analytically. Computer models can only offer numerical approximations, whereby the globe is divided into (digital) grid cells and the equations are solved approximately in discrete time steps. This makes it even theoretically impossible – that is, if one was in a position in which one could integrate every single process of the climate system into the model – to develop a computer model that creates a perfectly digital copy of the climate system.

On the other hand, from a practical point of view, not all parts of the climate system can structurally be represented equally well within a model. When developing a climate model, scientists have to make concessions, in one way or another, regarding which processes are (better) represented and which are not. Idealisations are a necessary part of any climate model. This is not a new development in science. After all, our common definition of model usually implies that it is an idealisation of something (Cartwright, 1983).[77] This is usually not considered a deficit but a clear advantage of models. Whether or not specific idealisations of a model are an asset depends on the particular purpose. For certain (e.g., educational) purposes, the Bohr model of the atom, which envisions the electron to circle the nucleus in a perfect circle, is sufficient. However, when one discusses more complex atom structures than the hydrogen atoms, for instance, to explain the Zeeman effect, one has to make use of other more advanced models. Still, for other purposes atoms need not be considered as more than a point charge, as in the kinetic theory of gases. This does not just hold for the use of models in physics. For example, ecological or economic models are well known for simplifying and reducing very complex systems.

The difference between climate models or many other kinds of computer simulations that deal with added epistemic challenges due to a high complex-

[76] Strictly speaking, both aspects, numerical approximation and structural uncertainty, actually do represent two distinct kinds of uncertainties, which, nevertheless, are so interconnected that in practice they are difficult to handle separately (Winsberg, 2018, p. 91).

[77] Model *in this context* should be understood in a broad sense (as Giere, 2006 for instance, does), not as a physical object or computer representation of a system but as a specific representation of a more abstract theory.

ity in the target system (see Chapter 2.1) and the examples above is the increasing difficulty to pin down exactly what effects these idealisations have on the model output. What one has to keep in mind is that a 'good' fit with observations cannot necessarily be traced back to a 'truthful' representation of the climate system by the model (see Chapter 3.2.4). It might very well that the good fit is due to compensating effects within the model, following from characteristic features of the model building process such as the fuzzy modularity and the tuning of models. As shown in Chapter 2.1, when assessing the 'quality' of a climate model, scientists face (at least) serious obstacles in gaining "analytical understanding" (Lenhard and Winsberg, 2010, p. 254). This is an issue that also affects the possibility of the application of RA to climate modelling, as will be discussed further below.

All of this makes it harder to pinpoint errors within the model and, in turn, to assess the uncertainties relating to the model structure. One method, as we will see, that climate scientists rely on to explore the structural uncertainty are multi-model ensemble studies (see Chapter 3.3.3.3).

3.3.3.2 Parameter uncertainty

Many processes relevant for the climate system concerning for instance cloud formation, radiation or vegetation growth cannot be resolved directly in the models. Any kind of process taking place on a subgride scale can only be integrated into the model in the form of parametrisations (McFarlane, 2011).

It is often the case that there are different options of how a specific process can be parametrised. Depending on how well the underlying mechanical processes are understood and can be expressed in terms of physical laws, parametrisations can be primarily derived from these and can be akin to a small model within the model with some empirically acquired parameters or based mostly on observationally derived approximations. As with the creation of models themselves, there are usually different options how to parametrise specific processes. In practice, parameter uncertainty is intertwined with structural uncertainty because the choice of a parameter value is very much contingent upon both the resolution of the model and the general model structure. The size of the grid cells will influence optimal parameter values and whether a certain process even necessarily needs to be parametrised. On the other hand, the overall model structure can influence the choice of parameter, for the particular interdependencies within the model might change what the best parameter values is. This is why the process of creating parametrisation (as well as tuning) is sometimes liked to being an "art form" (Edwards, 1999, p. 445).[78]

Improvements in the resolution can reduce the necessity to parameterise certain processes. However, this usually goes hand in hand with a significant increase in demand of computing power. Many non-negligible climate processes take place on a scale of a few kilometres, meters or way below that, such as those concerning the cloud microphysics. Current state-of-the-art ESMs have a resolution that is still far from that kind of resolution (see Chapter 2.1).

Parametrisations are deeply entrenched in the models. The consequence is that parametrisation schemes are not necessarily replaced or improved within a working model, even when there is a 'better' alternative available. As exchanging parametrisations require careful adjustments and tuning, doing so can be a costly undertaking (Guillemot, 2017).

However, it is not always computing power which limits scientists' abilities to resolve all relevant processes. Certain, specifically small-scale, processes are often not understood well enough to be resolved even if the scale of the models were small enough. As Knutti puts it: "There is simply no fundamental equation to describe how a tree grows" (Knutti, 2018, p. 328).

3.3.3.3 Second-order uncertainty

Why is it so complicated to pinpoint these uncertainties? Could one not just simply compare the model to observations? But as we have seen in Chapter 3.2.3, the available observational datasets themselves are generally extensively processed and come with a variety of uncertainties. Furthermore, one cannot simply interpolate from today's climate to that of the future under anthropogenic forcing. Structural epistemic obstacles such as not fully understood feedback processes make this impossible. Therefore, understanding the models, their strength, their shortcomings and, specifically for the purpose of climate-change projections, how all of this manifests in uncertainty estimates is paramount.

[78] A more in-depth discussion of the application of the term *art* to certain methods in climate modelling will following in Chapter 4.2.2.

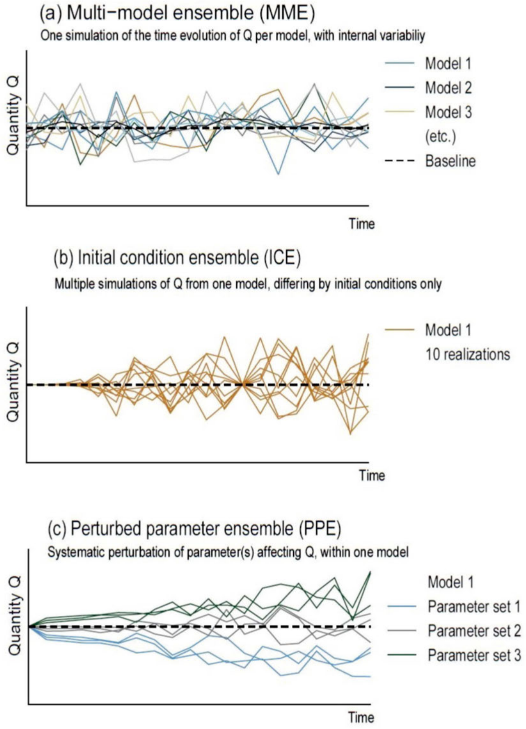
3.3.3.3.1 Ensemble studies

One way to explore these uncertainties are so-called ensemble studies (see Figure 5). To investigate uncertainties in the model structure, scientists compare different models under a fixed emission scenario with each other in so-called *multi-model ensemble studies* (MIPs). There is variety of model-intercomparison projects for all types of models. The most well-known among these is the Coupled Model Intercomparison Project (CMIP), which has been running since 1995 and is now in its 6th phase. The CMIP assessment of climate model uncertainty is an important contribution to the IPCC assessment reports. The uncertainty of parameter values can be investigated in a similar fashion. In *perturbed physics ensemble studies* (PPE) parameter values are varied in a model within the realm of what scientists consider reasonable to assess how this affects the overall model performance. A prominent example of such a project is done by *climateprediction.net*. In this project scientists have engaged the public so that they offer up free computing power on their private computers to run different versions of the same model (Stainforth et al., 2005). A newer development are *Initial Condition Ensembles* (ICE), which are ensembles based on one model with varying initial states under an otherwise fixed scenario, used to explore simulated internal variability (Chen et al., 2021, p. 222).

Particular MIPs have attracted the attention of philosophers of science, as they at first glance might seem similar to conventional statistical sampling methods but are not. For one, this would usually require sampling independently from the whole space of possible models. But as others (Parker, 2010; Winsberg, 2012) have pointed out, it is hard to imagine how one would even go about doing so. What is more, one would not even want to sample from the whole space of possible models. Introducing models of which we know that they are unrealistic into the uncertainty estimation seems to be contradictory to the purpose (Winsberg, 2012).

In addition climate models are also not truly *independent* of each other in the way it would commonly be required for statistical analysis. As we have already seen in Chapter 3.1.3.2, climate models have shared histories. Climate models are generally not entirely built from scratch but are usually at their core related to other models and contain parts (everything from lines of code to whole parametrisations schemes) also used in other models (Boé, 2018; Knutti et al., 2013).

Figure 5: Illustration of common types of model ensemble, simulating the time evolution of a quantity Q (such as global mean surface temperature)[79]



Source: Chen et al., 2021, p. 222, Figure 1.21

[79] (a) Multi-model ensemble, where each model has its own realization of the processes affecting Q, and its own internal variability around the baseline value (dashed line). The multi-model mean (black) is commonly taken as the ensemble average. (b) Initial condition ensemble, where several realizations from a single model are compared. These differ only by minute ('micro') perturbations to the initial conditions of the simulation, such that over time, internal variability will progress differently in each ensemble member. (c) Perturbed physics ensemble, which also compares realizations from a single model, but where one or more internal parameters that may affect the simulations of Q are systematically changed to allow for a quantification of the impact of those quantities on the model results. Additionally, each parameter set may be taken as the starting point for an initial condition ensemble. In this figure, each set has three ensemble members.

While MIPs might give the impression that it is an 'objective' approach to quantifying uncertainties, as Winsberg points out (2018, pp. 96–100), it is not an objective method in the sense of complete independence from the expert judgement of the scientists. Therefore, objective here essentially cannot mean anymore than 'not done by hand'.[80]

Traditionally, it is assumed that in an ensemble all models are equal and therefore are weighted the same. But considering that, one the one hand, models are often related to one another and, on the other hand, not all models are equally good at representing all aspects of the climate equally well, the assumption that all models in an ensemble should get the same 'vote' is questionable. Some attempts have indeed been made to weigh models according to performance and independence (e.g., Knutti et al., 2017), but there is no clear consensus among climate scientists on how this ought to be done (Chen et al., 2021, p. 226).

Further, philosophers and sociologists of science have in the past been attesting scientists a "herd mentality" (Winsberg, 2012; see also Sundberg, 2011) when it comes to constructing and evaluating climate models.[81] As the model development is not fully epistemically constricted, it is not uncommon that new models are matched to those that are well established. Social structures within the scientific community work in such a way that modelling groups usually try to avoid being the 'odd one' with a model standing out from the mass. Mikaela Sundberg (2011) argues model agreement creates a kind of "social authority" that scientists follow and adjust new models to in order to be taken seriously.[82]

'Subjective' estimates about the quality of the ensemble study in order to gain a full picture of all uncertainties are a necessary feature of climate-model assessment (Parker, 2014; Winsberg, 2018, pp. 96–102).[83] Nevertheless, ensem-

[80] Note that understanding objectivity in this way is similar to a definition of objectivity that is actually widespread in climate science, e.g., to describe specific algorithms (see Chapter 2.3 and Chapter 3.4.3).

[81] Winsberg actually compares this to Walter A. Shewhart's assessment of historic speed of light measurements, which converged despite being far from the right value (2012, p. 100, see also Shewhart and Deming, 1939).

[82] Sundberg argues that this kind of "social authority" does not just affect MIPs in climate science but also intercomparison projects in other fields of science that heavily rely on complex computer simulations such as astrophysics (2011).

[83] There is of course also the risk of a kind of herd mentality in more general terms when it comes to expert judgements. The IPCC remarks on this in the *Guidance Note for Lead*

ble studies are considered to be an essential element in exploring structural and parameter uncertainty in climate models and to investigate the effects of climate change. We will come back to this further below.

3.3.3.3.2 The quantification problem

Despite of all these uncertainties, policy makers often expect climate scientists to give clear and precise estimates of how the climate will change under what conditions. This is understandable; after all, in order to tackle the problem of climate change; it is useful to have as clear an understanding as possible of who and what is affected to what extent.

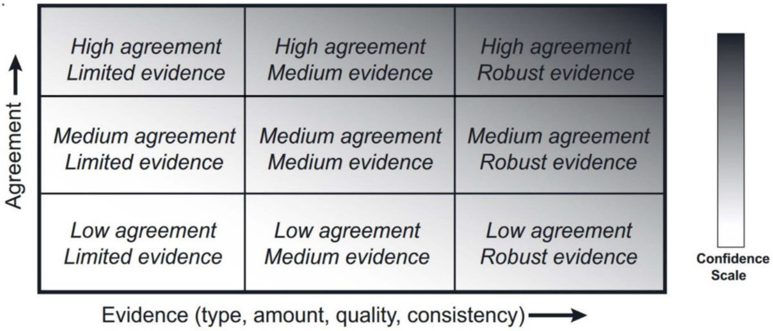
This, however, can potentially put scientists in a difficult situation when they try to stay true to their assessment of uncertainties and still convey helpful and concrete uncertainty estimates. But not providing any uncertainty estimates would not be feasible either as it bears the risk that someone less qualified might feel called to fill the gap.

To tackle this issue the Intergovernmental Panel on Climate Change (IPCC) has created a framework for its authors instructing them on how to communicate uncertainties. The *Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties*, which functioned as a baseline for both AR5 and AR6 (Chen et al., 2021, p. 169), gives the authors two options on how to convey the certainty or uncertainty of their findings:

> The AR5 will rely on two metrics for communicating the degree of certainty in key findings:
> · Confidence in the validity of a finding, based on the type, amount, quality, and consistency of evidence (e.g., mechanistic understanding, theory, data, models, expert judgment) and the degree of agreement. Confidence is expressed qualitatively.
> · Quantified measures of uncertainty in a finding expressed probabilistically (based on statistical analysis of observations or model results, or expert judgment). (Mastrandrea et al., 2010, p. 1)

Authors where the authors are advised to be "aware of a tendency for a group to converge on an expressed view and become overconfident in it" (Mastrandrea et al., 2010, p. 2).

Figure 6: A depiction of evidence and agreement statements and their relationship to confidence. Confidence increases towards the top-right corner as suggested by the increasing strength of shading.



High agreement Limited evidence	High agreement Medium evidence	High agreement Robust evidence
Medium agreement Limited evidence	Medium agreement Medium evidence	Medium agreement Robust evidence
Low agreement Limited evidence	Low agreement Medium evidence	Low agreement Robust evidence

Source: Mastrandrea et al., 2010, p. 3, Figure 1

The IPCC accompanies this with two helpful charts, which further demonstrate how confidence and likelihood should to be understood and communicated (Figure 6 and the likelihood scale already introduced in Chapter 3.1.3.3 Figure 2).

On the one hand, scientists can express confidence in a qualitative way as relating to evidence and agreement. It is stressed in the Guidance Notes that *confidence* should not be interpreted probabilistically. On the other hand, the authors are encouraged when they come to the conclusion that evidence and confidence is sufficient to articulate assumptions about the certainty of specific events or results as *likelihood*. Here, specific terms to convey findings of likelihood (such as *Very Likely*, *Likely*, *Unlikely*) are assigned to margins of outcome probability (90–100 %, 66–100 %, 0–33 %), as shown in Figure 2.

Therefore, the authors of the IPCC assessment report are given a framework that is at the same time calibrated but also somewhat flexible to account for the specificities of a particular hypothesis and the evidence for it. Nevertheless, the *Guidance Note for Lead Authors* also advices the authors of the IPCC report to consider potential pitfalls in how the wording of their findings might be misinterpreted:

> Be aware that the way in which a statement is framed will have an effect on how it is interpreted (e.g., a 10% chance of dying is interpreted more negatively than a 90% chance of surviving). Consider reciprocal statements to

> avoid value-laden interpretations (e.g., report chances both of dying and of surviving). (Mastrandrea et al., 2010)

The IPCC as the UN's organisation that is entrusted with gathering and assessing the current state of climate science is, as Edwards puts it, a unique "hybrid scientific/political organization" (1999, p. 460). Therefore, it has to be particular careful when it comes to communication uncertainties as the reports are not just read by members of the scientific community who are familiar with the conventions but also outsiders who might misread those writings.

However, despite this strict framework, several studies have shown that assessments made according to these guidelines are often misunderstood by laypersons. Particularly wider uncertainty intervals are often misinterpreted to mean that the scientists were less certain than when the interval was smaller (Løhre et al., 2019). How the scientists' assessments are interpreted can also be dependent upon cultural backgrounds (Harris et al., 2013). While AR6 still makes use of the framework given in the Guidance Note originally developed for AR5, the difficulties with communicating uncertainties is also acknowledged there (Chen et al., 2021, p. 171).

After having established, first, the different sources of uncertainty that come along with climate modelling, second, the difficulties in determining how strong these uncertainties are and, third, the language the IPCC applies in communication these uncertainties, the question now remains how the IPCC comes to conclusions, despite all these uncertainties, about the likelihood of (and confidence in) hypotheses about the various effects of increased climate forcing. In order to discuss this question, we will return to RA.

3.3.3.4 Robustness revisited

Returning to RA, the central questions now are: what should be inferred from the fact that projections from different climate models agree? And can we tell if the models agree because of some common essential and true core of the models or because of some specific idealisations of the models? As discussed above, ensembles are not statistical sampling methods. Philosophers and climate scientists generally agree that due to the lack of interdependence of models ensembles cannot be regarded as statistical sampling methods (e.g., Knutti et al., 2017; Parker, 2018; Winsberg, 2018). But is there still some kind of epistemic significance to ensembles of models? The follow-up to this question is: can *more* models increase confidence in their output?

Some philosophers have indeed expressed hope that, in a certain respect, RA might be of help here (Lloyd, 2010, 2009; Winsberg, 2018), while others have been less convinced (Parker, 2011).

In the context of climate modelling RA was first discussed by Elisabeth Lloyd (2009). She argues that Weisberg's (2006) version of model robustness, as represented in Chapter 3.3.2, could also be applied to climate model ensembles. One might, for instance, find in respect to an ensemble of models that

> in all of them [the models] there is a significant role played by greenhouse gases in the late twentieth-century warming of the global climate, and that these are linked to the surface temperature rising in the equations, despite the fact that climate models vary in their assumptions about other aspects of climate. Thus, we would have an analysis isolating greenhouse gases linked to temperature rise (the common structure), and a robust theorem linking greenhouse gases to the robust property, the outcome of rising global mean temperature. (Lloyd, 2009, p. 220)[84]

However, even if all models show a connection between rising temperatures and greenhouse gases, the question still remains how can we be certain that greenhouse gas emission is the relevant factor. To answer this, Lloyd notes, Weisberg makes an "implicit appeal" to a variety of evidence argument:

> he is explicitly appealing to a range of instances of fit of the model over different parameter values, parameter space, or laws. It is against this background of differing model constructions that the core structure occurs and causes the robust property to appear, and it is the degree of this variety of fit for which the model has been verified that determines how confident we should be in the causal connection. (Lloyd, 2010, p. 981)

Lloyd concludes that when a diverse set of models agree so that a robust theorem can be formulated and they also show other instances of fit, we have good reason to be confident in regards to the robust property. Besides a good fit with observations of GMST of the 20th century, different climate models, even

[84] The robustness theorem then would be: "Ceteris paribus, if [Greenhouse gases relate in lawlike interaction with the energy budget of the earth] obtains, then [increased global mean temperature] will obtain" (Lloyd, 2010, p. 950).

of a specific type, contain a variety of different background assumptions,[85] parametrisations and parameter values which are, while sometimes contradictory, in themselves empirically supported, Lloyd argues (see also Lloyd 2015). Thus, Lloyd contends that model robustness in this context can have a confirmatory, not just heuristic dimension when not only multiple models with the same causal core converge towards a specific result, but when there are also a variety of diverse but empirically supported modelling assumptions.

Wendy Parker (2011), by contrast, argues that ensembles of climate models on their own do not warrant any conclusions about the truthfulness of or confidence in an ensemble result. Building on Orzack and Sober's argument (1993), Parker reconstructs RA for an ensemble of climate models as follows:

> 1. It is likely that one of the models in this collection is true.
> 2. Each of the models in this collection logically entails hypothesis *H*.
> It is likely that *H*. (Parker, 2011, p. 583)

Parker notes that such an argument is problematic in the context of scientific models, as idealisations and simplifications are an unavoidable feature. In fact, idealisations are what make models models. Thus, in some way, a model is always false.

However, one can transform the argument in such a way that only the "likely adequacy" (Parker, 2011, p. 584) of a model for a specific purpose is required:

> 1'. It is likely that at least one simulation in this collection is indicating correctly regarding hypothesis *H*.
> 2'. Each of the simulations in this collection indicates the truth of *H*.
> It is likely that *H*. (Parker, 2011, p. 584)

However, Parker argues that today's climate models neither in respect to the ensemble's performance nor its construction can fulfil the likely-adequacy condition. Concerning the latter, Parker points out that (as already discussed above) ensembles do not sample from the whole space of possible model but are *ensembles of opportunities*. Therefore, it cannot be argued that it is likely that one of the models in an ensemble is indicating correctly regarding *H*

[85] Lloyd argues that even models with limited changes in the model structure can be considered sufficiently diverse in the case of climate models, as the non-linearity and feedback process will sufficiently diversify the model output (2015, p. 65)

as the ensemble does not reflect the whole range uncertainty. And as far as model performance is concerned, Parker is also sceptical that a good fit with observations in respect to a particular variable can be rightfully interpreted as an indication that it is likely that one model in the ensemble comes close to predicting the true value of that variable at some point in the future because of the intricate ways of model-data interdependency (see Chapter 3.2).[86]

Eric Winsberg (2018, p. 179) interprets the reason that Lloyd and Parker come to such different conclusions about the applicability of RA to climate science is that they essentially ask two different questions. Lloyd's goal, according to Winsberg, is to explore whether models plus other evidence could support a climate science hypothesis, whereas Parker's approach to RA focuses only on models. Considering these different premises, Winsberg comes to the conclusion, as we will see in the following, that both Parker and Lloyd make valid points concerning RA.

Winsberg begins his analysis of RA (2018, pp. 183–206) by expanding the notion of RA beyond models to a variety of types of evidence coming from a combination of different sources, such as experiments and observations. Further, stemming from the question what is actually meant by "sufficiently diverse" (Weisberg, 2006) for a set of models or/and other evidence to be considered for RA, Winsberg sets out to look for a concept of *RA-diversity* that also "acknowledges that science at best offers grounds for increasing one's degree of belief in a hypothesis" (2018, p. 185). Inspired by Schupach (2018) Winsberg argues that as all climate models share at least some common assumptions (and one would hope so), they cannot be considered to be (fully) independent of one another, one would need a concept of RA-diversity that is not, as is commonly assumed, built on a notion of "probabilistic independence". That is, one would need to find a kind of RA-diversity that also holds for those cases where there is some kind of entanglement either among the pieces of evidence or methodology. Jonah Schupbach (2018) provides such concept of RA-diversity. Studying the application of RA in the context of different scientific practices such

[86] Parker further explores whether "increased confidence in H is warranted" or "the security of a claim to have evidence for H is enhanced" (Parker, 2011, p. 581), instead of "likely adequacy", might be more successful approaches. But for similar reasons to those already discussed, Parker argues that these weaker requirements do not suffice to consider RA successful in the case of climate model ensemble studies.

as experiments, observations and models (computer and other kinds), Schupbach asks the question "what is accomplished in successful RAs by introducing diverse means of detection" (2018, pp. 286–287). He answers this question by arguing that these diverse detection methods provide ideally "competing explanations". He proposes a concept of RA-diversity that is characterised by increased confidence in a hypothesis when a new piece of evidence is added. It is defined in the following way:

> Explanatory RA-diversity: Means of detecting *R* are RA diverse with respect to potential explanations (target hypothesis) *H* and its competitors to the extent that their detection (R_1, R_2, …, R_n) can be put into a sequence for which any member is explanatory discriminating between *H* and some competing explanation(s) not yet ruled out by the prior members of that sequence. (Schupbach, 2018, p. 288)

Based on this, Winsberg calls it the "cumulative epistemic power" (2018, p. 185) of a set of models or other types of evidence when it is rational to assume that every new piece of evidence increases our confidence in a hypothesis.

Appling this concept of RA-diversity to climate modelling, as Winsberg does (2018, pp. 192–193), one might assume a model that 'detects' that equilibrium climate sensitivity (ECS), that is, in short the change of the global mean surface temperature after a doubling of CO_2 in the atmosphere, is above 2 °C. However, the question arises if there are no other explanation for this model result other than that the hypothesis is correct such as some kind of distortion in the model, for example, the particular grid size or the wrong cloud parametrisation. Thus, one has to ask what other detection procedures exist that could rule out these other competing explanations, e.g., by trying out an ensemble of models with different grid sizes or other cloud parametrisations. This process, as Winsberg points out, requires making specific case-dependent decisions about whether or not the competing hypotheses are sufficiently dismissed.

This also means, Winsberg notes, that, whether a set of models is *adequately RA-diverse*, can only be determined with respect to a specific hypothesis and with respect to the question asked and to what extent one can rule out competing explanations. Nor would saying that a set of models and/or other lines of evidence is RA-diverse necessarily be a statement on whether or not to accept a hypothesis. It is only a sign for cumulative epistemic power, concludes Winsberg, that is, being on the right track.[87]

Though RA defined in this way is applicable to all sorts of lines of evidence, it may also be applied to just models. Whether or not an ensemble of models is considered RA-diverse is only a question of the particular hypothesis, argues Winsberg: "You can make an ensemble of opportunity RA diverse without altering the ensemble by altering the hypothesis" (Winsberg, 2018, p. 202).

Further, establishing explanatory RA-diversity will frequently require an understanding of the underlying structure of the model and the climate system in general with respect to the relevant processes, as we will see in the following.[88]

As an example of how this definition of RA-reasoning can be applied to climate-change hypotheses, Winsberg examines in a case study the problems of establishing a value for ECS (2018, pp. 194–206). He particularly focuses on an estimate from CMPI5 of ECS being between 2.1 °C and 4.7 °C, as reported in AR5 (Flato et al., 2013, p. 818).[89] The difficulty in estimating ECS can be attributed to the fact that it is not just the outcome of increased CO_2 in the atmosphere but also several feedback processes, with cloud feedbacks considered to be the biggest issue.

Feedbacks are also at the root of why Winsberg is sceptical that ensembles of opportunity (on their own) can provide robust evidence, that is, it cannot be demonstrated that the ensemble is sufficiently RA-diverse when it comes an estimate of ECS in the sense of the hypothesis above. Let's see why. First of all,

[87] It is important to stress that *RA-diverse* is different from *sufficiently diverse* insofar as RA-diverse does only imply that a model set "gets better as it gets larger" (Winsberg, 2018, p. 186). Whether or not to accept a specific hypothesis is, then, context-dependent and among other things a question of inductive risks (see Chapter 3.1).

[88] Winsberg links this to Knutti's definition of process understanding (Knutti, 2018, see also Chapter 3.2.4). In fact, Winsberg claims "gaining process understanding is not necessarily a separate kind of epistemic activity from RA, and the two are complementary, rather than competing, accounts of how we gain confidence in model results" (Winsberg, 2018, p. 202).

[89] While the latest IPCC report (AR6) does no longer directly consider models for assessing ECS, it is nevertheless helpful for the questions asked here, that is, how RA can be applied to models as well as other sources of evidence, to take a look at Winsberg's appraisal of the reasoning process behind the AR5 assessment of ECS, even if it is somewhat out-dated.

taking into account all we know about the challenges of creating ESM, all models agreeing on a particular estimated range of ECS could still be an "artefact of the systematic failure of all the models to accurately capture all of the feedbacks – with cloud feedbacks being an especially likely candidate" (Winsberg, 2018, p. 196). Further, even if the models fit well with the available observational data of the relevant feedback process, this does not necessarily mean – as we have seen in Chapter 3.2.4 – that the models would also necessarily adequately capture the effect of feedbacks for a possible future climate that differs significantly from what we currently have observations for. It might as well be that the models only fit so well with the available data because of some compensating errors in the models that will cease to compensate in the same way under future climate change conditions. Thus, considering the significance of cloud feedbacks to assessing ECS, Winsberg concludes that for the models of an ensemble to be RA-diverse regarding a hypothesis about ECS, they would have to not just accurately model the cloud feedbacks with respect to the observable past; one would also have to eliminate the possibility of the model fitting well with the data due to error compensation. Only if this were accomplished, argues Winsberg, then "there is a high probability that we are correctly detecting a hypothesis about cloud feedback in the future climate" (2018, p. 197).

In this context Winsberg also points out a more recent trend in climate-change assessment to tackle ECS and similar problems: trying to find so called *emergent constraints* (Chen et al., 2021, p. 225; Winsberg, 2018, pp. 197–201). The goal of this method is to reduce uncertainties in climate-change projections by establishing a relationship between a future climate-change response and present-day observations. Let us assume we are interested in a variable *b* (also called the predictand), e.g., the intensity of a specific feedback process that models in the ensemble projection do not agree on. However, we suspect that there is a correlating relationship between variable *b* and another variable *a* (also called predictor) of a process that is taking place on timescale for which there are good observations. Such a relationship might, for instance, be the snow-albedo feedback (that is, the increase in absorption of solar radiation due to the increase of ice melting because the surface is warming) which is taking place both in a seasonal cycle and under (longer term) forcing conditions (Hall and Qu, 2006). This relationship is then established by running the ensemble once over a short time period to determine the model spread of variable *a* and once under a long-term forcing scenario to determine the model spread of variable *b*. The model output for both variables is plotted, with variable *a* on the x-axis and variable *b* on the y-axis, and every point on the graph representing one model (see Figure 7).[90] An *emergent constraint* is considered to be "trustworthy" (Caldwell et al., 2018) or "confirmed" (Hall et al., 2019) when it can be further argued that the mechanical process behind the correlation of *a* and *b* is well understood and the emergent constraint is tested 'out of sample' with an ensemble that was not used in finding the emergent constraint, so as to rule out that the emergent relationship is not just a coincidence, as a result of compensating errors and lack of diversity of models in the ensemble (Hall et al., 2019).[91]

Winsberg argues that the reasoning process behind emergent constraints can be considered "one of the best RA reasoning […] in climate science" (2018, p. 197) to the extent that it is a way to systematically rule out alternative explanations for a hypothesis by establishing and testing the stability of the particular underlying process (2018, p. 201).

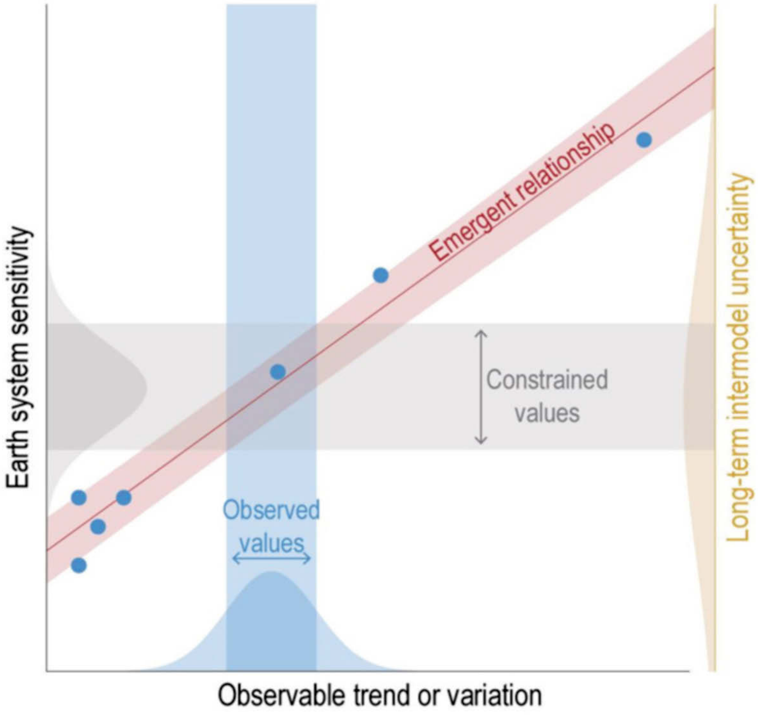
While this has been done successfully, e.g., in the case of the snow-albedo feedback (Hall and Qu, 2006), this kind of reasoning is much more difficult to establish when it comes to ECS, which is affected by many different kinds of feedback processes. Thus, returning back to the question if an ensemble of models can be considered explanatory RA-diverse concerning a hypothesis about ECS, Winsberg argues that this requires that it is first demonstrated sufficiently that the ensemble is exploratory RA-diverse in the sense of an emergent-constraint reasoning process in respect to every single feedback effect that gives rise to ECS.

That is why Winsberg is sceptical that a robust reasoning concerning a hypothesis about ECS could be arrived at based on an ensemble of models alone (2018, p. 199). However, Winsberg's definition of explanatory RA is not restricted to models. Thus, if one takes into account that scientists also have access to other detection methods in the form of instrumental records and proxy paleoclimate data as independent lines of evidence then one can see how the authors of AR5 nevertheless come to the following conclusion:

[90] This, of course, requires that the range of observations is within the interval of the model spread for variable *a*, so that variable *b* can actually be constrained.

[91] If all models in the ensemble are related to each other too much, then the suspicion that the emergent relationship might be an artefact of model error is more pressing than when the models are more independent (Brient, 2020).

Figure 7: The principle of emergent constraints.[92]



Source: Chen et al., 2021, p. 225, Figure 1.23

Based on the combined evidence from observed climate change including the observed 20th century warming, climate models, feed-back analysis and

[92] An ensemble of models (blue dots) defines a relationship between an observable mean, trend or variation in the climate (x-axis) and an uncertain projection, climate sensitivity or feedback (y-axis). An observation of the x-axis variable can then be combined with the model-derived relationship to provide a tighter estimate of the climate projection, sensitivity or feedback on the y-axis. Figure adapted from Eyring et al. (2019).

> paleoclimate, ECS is *likely* in the range 1.5°C to 4.5°C with *high confidence*. The combined evidence increases the confidence in this final assessment compared to that based on the observed warming and paleoclimate only. ECS is positive, *extremely unlikely* less than 1°C (*high confidence*), and *very unlikely* greater than 6°C (*medium confidence*). (Collins et al., 2013, p. 1111)

This statement is based on a set of detection methods that is RA diverse, Winsberg argues, for the following reason (2018, pp. 202–206): models, instrumental records and paeleoclimate data applied as detection methods for ECS all come along with specific uncertainties. However, if not all lines of evidence are susceptible to the same uncertainties (or at least not to the same degree) and one treats "each of these sources of uncertainty […] as alternative possible explanations of various hypotheses detections" (Winsberg, 2018, p. 205), then it is possible to see how scientists can rule out different alternative explanations one after another and provide robust evidence for a likely range of ECS as well as the possible upper and lower limits, Winsberg concludes.

For instance, paleoclimate data is not much prone to errors coming from internal variations but most probably will suffer from higher measurement uncertainty and might rest on a different base state as many millions of years ago the climate might have been quite different to our current. The latter does not affect instrumental records and when the hypothesis is broad enough (as, e.g., that ECS is between 1.5 °C and 4.5 °C, see above) measurement uncertainties are also less of a concern here. Both types of detection methods are also less susceptible to model errors, as simulations of the type of ESM are, and so on.[93] This kind of reasoning is at the heart of the statement about ECS in AR5, Winsberg argues.[94]

[93] Winsberg also notes that RA reasoning is also helpful to understand why it is so difficult to constrict ECS at the upper end. Considering the question if there are alternative explanations for why the current detection methods do not project a higher value, there could, for instance, be a yet unknown feedback process, which instrumental records do not yet detect. Winsberg concludes, that we "would probably only expect to see it in the millions-of-year-scale paleodata – but those data sets have enough uncertainty that they are poor at eliminating such a hypothesis" (2018, pp. 206), which makes constricting ECS at the upper end so difficult.

[94] For a more in-depth analysis of the argument of the applicability of the idea of RA to the reasoning about the value of ECS, see Winsberg (2018, pp. 203–206). He follows the scientific argument made in Knutti and Hegerl (2008).

In this context it is worthwhile to mention that compared to the assessment of AR5 the authors of AR6 actually do consider *emergent constraints*, alongside *instrumental records*, *paleoclimates* and *understanding of the climate processes* as lines of evidence for the following assessment of ECS:

> Based on multiple lines of evidence, the very likely range of equilibrium climate sensitivity is between 2°C (high confidence) and 5°C (medium confidence). The AR6 assessed best estimate is 3°C with a likely range of 2.5°C to 4°C (high confidence), compared to 1.5°C to 4.5°C in AR5, which did not provide a best estimate. (IPCC, 2021b)

With respect to the discussion above a few things are noteworthy here. First of all, while emergent constraints are now considered a line of evidence, they are not the only line of evidence.[95] The authors particularly point out that emergent constraints in AR6 are not combined "to provide very strong evidence on ECS" because there are still cross-dependencies between different emergent constraints and it is still too new a technique to rule out that there may not be unaccounted systematic biases (Forster et al., 2021, p. 1005).

Secondly, models are no longer a direct line of evidence and find their way into the assessment only indirectly (for instance, through emergent constraints and because they inform process understanding of feedbacks). The authors cite issues pertaining to the specifics of climate-model construction discussed in this chapter, such as lack of model independency, analytical intractability and the difficulties of evaluating and weighing models adequately as reasons for this (Forster et al., 2021, pp. 1007–1009).

Last but not least, the authors argue that one reason for the improvements in the assessment of ECS from AR5 to AR6 was the application of a new structured (Bayesian) approach of taking different lines of evidence[96] into account

[95] In respect to Winsberg's argument it is interesting to note that authors of AR6 differentiate between two kinds of emergent constraints on ECS: "(i) those that are based on global or near-global indices, such as global surface temperature and the TOA energy budget; and (ii) those that are more focussed on physical processes, such as the fidelity of phenomena related to low-level cloud feedbacks or present-day climate biases" (Forster et al., 2021, pp. 1004–1005). Only the first kind of emergent constraints are taken into account in the assessment of ECS with the authors citing concerns about possible biases.

[96] Sherwood at al. (2020) base their assessment only on three lines of evidence: instrumental record, paleoclimate data and process understanding. Emergent constraints are not considered as a distinct line of evidence.

as done by Sherwood et al. (2020). Sherwood et al. note that the broad agreement of lines of evidence works as "mutual reinforcement" (2020, pp. 73–74; see also Forster et al., 2021, p. 993). They apply what Stevens et al. (2016) have called a storyline-approach to combine different lines of evidence by laying "out all the circumstances that would have to hold for the climate sensitivity to be very low or high given all the evidence" (Sherwood et al., 2020, p. 2). This shows some similarities to Winsberg's concept of RA in the sense that it is not just an appraisal of models but different types of evidence and more importantly, that it is a systematic approach to rule out alternatives (though in this case alternative storylines, not alternative explanations for a hypothesis).[97]

3.3.4 Conclusion

There are, as we have seen in this chapter, considerable obstacles to be navigated in order for climate scientists to make any kind of assumption about the future of the climate. However, while the sources of uncertainty are manifold and often not easy to assess or minimise, this does not mean that the uncertainties are so overwhelming that no conclusion about the anthropogenic climate change can be made. Quite to the contrary, many of the essential variables and processes that determine how the climate changes are well understood. Take, for example, the case of ECS discussed in this chapter. While the estimate of the range has been refined since the assessment in the famous so called *Charney Report* in 1979, the estimate that it lies between 1.5 °C and 4 °C has been consistently confirmed (Charney et al., 1979; Forster et al., 2021, pp. 1006–1007).

In public debates about climate change it has often been argued that the models are not good enough, that the models disagree with the data or that there is still to much uncertainty. But a central insight from Winsberg's analysis of the applicability of RA in climate science that I think is important to highlight here is that the question of whether or not an ensemble of models is robust is the wrong question to ask. Instead the question ought to be whether or not a specific hypothesis can be supported or refuted with the help of very hypothesis-specific detection methods. More broadly speaking, the question

[97] For another similarity note that concerning the unavoidability of expert judgement in the whole process Sherwood et al. also point out "solid *qualitative* understanding of how the evidence stacks up is at least as important as any probabilities we assign" (2020, p. 73).

whether or not there are good reasons for accepting or rejecting a specific hypothesis about climate change is usually not a question of model or data or even models versus data but a elaborated reasoning process based on a combination of different lines of evidence.

The argument that has been made here is that, despite the complexity of the climate system and all epistemic challenges, many important questions concerning the future of the climate can be dealt with with careful reasoning. But what climate science (and science in general, for that matter) cannot provide are fully uncertainty-free answers to every question in an irrefutable way and without any doubt. This sometimes clash with public expectation. However, when scientists are faced with demands to only make yes-or-no-statements giving in to these demands does not just misrepresent the scale of scientific knowledge accumulated but also bears the risk of delaying taking action to mitigate climate change, as Isaac points out:

> Typically, policy-relevant issues are publicly discussed in binary terms […] yet the relevant science is more appropriately framed in terms of degrees of certainty or evidential support […] A public rhetoric of bivalence obscures the nature of the scientific contribution to our knowledge of the world and undermines its effective use in policy choice. (Isaac, 2014, p. 43)

In public discourse the demand for 'better' science is often framed as a necessity in order to assess whether or not costly mitigation measures actually have to be taken. In practice, though, this has often been misused as a stalling tactic by particular interest groups to advocate against policies which would be to their disadvantage (Howe, 2014; Oreskes and Conway, 2010). Heymann et al. (2017b) interpret this attitude towards uncertainties in science as a consequence of a view of science which emphasises prediction making as a major trait of science:

> As a consequence of their political, cultural, and economic status and value, tremendous resources flow towards the establishment and operation of cultures of prediction. These investments do not always serve the support or justification of decision making and politics, but can also serve to delay or replace decision making and politics – particularly in the case of contested issues with strong inherent political risks. A commonplace argument for the replacement of effective politics is the call for further research, for example, due to perceived or alleged uncertainty and the lack of sufficient knowledge apparently required to make strident decisions. (Heymann et al., 2017b, p. 8)

Thus, the ‘sound science’-argument can quickly turn into a ‘throwing out the baby with the bathwater’ type of reasoning as it neglects the many questions which can be answered with a broad consensus among climate scientists: chiefly amongst them that anthropogenic climate change is happening. But the knowledge that scientists have accumulated about climate change goes way beyond that. Nevertheless, estimating uncertainties is an intricate process where many assumptions have to be made and lines of evidence have to be weighed and examined requiring carful reasoning.[98]

3.4 Looking back and a tentative look forward

This chapter has shown why three specific widespread ideals about how science does and should operate are inadequate to describe actual scientific practice, particularly when there are, like in climate science, additional epistemic challenges due to the high complexity of the target system. All three ideals have a history of being upheld as signs for good, reliable, adequate science. Science being, firstly, a value-free enterprise that, secondly, works with theories which are easily and unambiguously assessable with the help of experiments or observations and, thirdly, that provides clear binary predictions about the future is commonly considered to be hallmarks of good science. In the past, following these ideals has often been considered to be what makes science special and distinguishes science from other human endeavours.

In the context of science dealing with highly complex systems it becomes apparent what has always been the case: science cannot, has never and, most importantly, does not have to live up to these ideals. But what does the inadequacy of these ideals mean for our ability, specifically from the perspective of an outsider to the scientific community, to draw conclusions about the quality or credibility of the methods and hypotheses of, for example, climate science? Is there another way to ground our trust in science? Fortunately, I think there is. It requires shifting the focus from the characteristics of either the individual

[98] The IPCC here functions specifically as an organisation that screens and assesses all existing research. It also provides policymakers and the general public with summaries and estimates of uncertainties. It should be pointed out that the IPCC is an unusual institution that has only few equivalents in other sciences. The existence of the IPCC is not just a sign of the relevance of climate research for society but also of the complexity of the system as well as the overall issue.

scientist or the research on to the social structures of science. This reasoning and what follows from this will be further explored in the next chapter.

Here, however, I want to first take a quick look back. In Chapter 2, I have introduced three different 'recurring themes' as preliminary remarks: the *epistemic challenges of highly complex systems*, the *DJ distinction* and *scientific objectivity*. I stated that theses concepts would reappear in the in our examination of the three ideals. Now I will revisit these concepts and see how they pan out, specifically in the context of climate science. We will see that these themes were not just contributing factors to the development and implementation of the ideals and that they are a turn of conflict for these ideals with actual science practice but also that the issues with these concepts will also give a first indication of a way out of this dilemma.

3.4.1 Complexity and understanding

Chapter 2.1 gave an introduction to the complexity of the climate system and how scientists navigate this complexity with help of computer simulations. One concern which was raised there is that modelers will inevitably have to deal with some degree of "epistemic opacity" (Baumberger et al., 2017; Humphrey, 2004) or "analytical intractability" (Lenhard and Winsberg, 2010) due to the complexity of the system and models. One question that follows from this is what this means for our ability to achieve *understanding*. While it has traditionally been a primary aim and motivation for scientists to understand 'why the things are the way they are' or 'how things work', Johannes Lenhard (2020) argues that computer simulations have shifted our perception of what science can and cannot accomplish in terms of acquiring *understanding*. He comes to the conclusion that, while complex climate models, like ESMs, are employed by climate scientists to explore and gain knowledge about the global climate system, they also contribute to a reduction in access to understanding due to their high complexity. On the one hand, many essential aspects of the model performance derive from the physical principles and fundamental equations that the model is based on. On the other hand, the models consist, of course, of much more than just these basic equations. ESMs display an intricate relationship between fundamental equations, semi-empirical parametrisations and tuning that is further intertwined through the iterative process of model construction and improvement (see Chapter 3.2.3.3). Because of this Lenhard argues that the conventional strategies for acquiring understanding in science by dealing with the complexity through "stripping off aspects until only the essence remains" (Lenhard, 2020, p. 2) cannot be applied. This is what Lenhard refers to as the "dilemma of growth" (2020, p. 2); While increasing complexity is a necessary feature of the models to live up to the high complexity of the climate system, the high complexity of the models limits the extent to which it is possible to reach the aim of "getting to the essence of the mechanism" (Manabe, 2006; as quoted in Lenhard, 2020, p. 2).[99]

Nevertheless, Lenhard observes that climate scientists often develop a "feeling" for the model and its behaviour. Scientists regularly rely on the experience they have with the model and derive some informal knowledge of how certain adjustments will most likely affect the performance of the model in question even though the exact inner-model processes that generate this model behaviour are not fully transparent. Inspired by Max Weber's concept of *verstehendes Erklären* (understanding-explanation (1913))[100], Lenhard argues that this is a way to circumvent what he calls elsewhere the "complexity barrier" (2019) preventing scientists from going the more established route to understanding:

[99] Lenhard notes that one possible strategy to deal with the "dilemma of growth" might turn out to be resorting to a 'hierarchy of models': "The hope is that small and well-understood models can be knitted together in larger hierarchies so that understanding extends to the whole. However, the prospects of this approach are not yet clear, in part because modularity tends to erode in larger simulation models, leading to a problem of 'holism'" (Lenhard, 2020, p. 2). In respect to the problem of finding a 'hierarchy of models' see also Held (2005).

[100] In this, as Lenhard argues, two different approaches to conceptualising *understanding* in philosophy of science are combined. Both centre on the terms *explanation* and *understanding* but define the relationship differently. On the one hand, there is the notion that understanding can be gained once an *explanation* in form of a (logical) *derivation* from basic principles is reached. On the other hand, there is the view – coming from a hermeneutical perspective – that sees explanation and understanding as belonging to two different research fields. Explanation is what can be achieved in science. Understanding, though, belongs to the humanities, where understanding is reached when something, or rather someone, behaves in a way that *matches* what one has anticipated.

Lenhard argues that both *derivation* and *match* can be found in Weber's conception of *verstehendes Erklären* and in the kind of understanding that is reached in climate modelling. The equivalent to *matching* in climate modelling is the "feeling" that Lenhard remarks climate scientists develop for their models, but to some degree, there is also an element of *derivation* because at its core the models are, despite semi-empirical parametrisations and tuning, still built on some fundamental equations.

> Such acquaintance with model behavior can be a work-around for building an adequate "inner" representation [...] when simplification/idealization strategies are *not* available—as in the case of ESM. However, this work-around does not lead to understanding in the traditional sense. There are no simple models involved that would enable understanding by "capturing the essence of a phenomenon" (Held, 2005, 1609). Nevertheless, simulation provides understanding—if only in the weaker, pragmatic sense of getting acquainted with model behavior. (Lenhard, 2020, p. 3)

Lenhard contrasts this with a more conventional conception of *explaining* and *understanding* that is based fully on "analytic derivation from first principles" (Lenhard, 2020, p. 3), which in the context of climate modelling (at least at this point)[101] cannot be accomplished. However, Lenhard argues, what can be achieved is a pragmatic conception of understanding that might function as a substitute so long as no other, more satisfying, conceptualisation of understanding can realised (see also van Fraassen, 1980).

3.4.2 Discovery and justification

Another concept introduced in Chapter 2.2 was the distinction between *context of discovery* and *context of justification*. The DJ-distinction was of relevance both for the emergence of the value-free ideal and the theory-centred view of philosophy of science.

Although the aftermath of the brief popularity of the distinction at the middle of the last century can still be felt (see Chapter 2.2), it is a conception that does not seem to have gained much traction in current philosophy of science. For example, in the edited collection *Revisiting Discovery and Justification* edited by Jutta Schickore and Friedrich Steinle (2006b), none of the authors argues to uphold the distinction as a dichotomy in any strong interpretation of the concept. On the contrary, many claim that a context distinction going beyond a weak form in the sense of Hoyningen-Huene's differentiation between a normative and descriptive view of science cannot be maintained (see Chapter 2.2). Specifically, the claim that justification is "'the other' or 'the opposite' of theory construction, experimentation or indeed discovery" is rejected by theses authors (Schickore and Steinle, 2006a, p. xiii). On the contrary, Schickore and

[101] It should be pointed out here that Lenhard notes that his argument is not that this concept of pragmatic understanding "should or in fact must become the goal" in climate science (2020, p. 3).

Steinle note in the introduction of the aforementioned book that "[d]iscovery, in any meaningful understanding of the concept, is a prolonged activity that involves both the generation and fixation of knowledge claims" (2006a, p. xiii). In this sense, e.g., Steinle (2006) and Arabatzis (2006) argue that discovery and justification in scientific practice go hand in hand. Steinle argues that this is, especially, the case in the event of *exploratory experiments*:

> Exploratory experimentation is concerned with developing regularities and appropriate concepts. If it is successful, this success consists in formulating ever more general laws. One may well ask whether such laws have then been discovered or justified: after all, in common language we often speak of laws as having been "discovered" by Galileo, Boyle, Hooke, or Mariotte, for example. As soon as we try to clarify our concepts, however, such talk immediately becomes inappropriate: at the moment when laws are formulated in the research process, they are discovered and justified at the same time. Even if a researcher had initially just a speculation of a possible empirical law, she would conceive this law as being "discovered" only in the moment when it was fully supported, i.e., justified. (Steinle, 2006, p. 187)

The notion that we usually only talk of something (such as a theory, phenomena, object) as being 'discovered', when the belief in it is also considered to be justified by the particular scientific community, is also stated by Arabatzis:

> A mere hypothesis to the effect that a new entity exists would not qualify as a discovery of that entity. The justification of that hypothesis would be a constitutive characteristic of that discovery. The context of discovery is "laden" with the context of justification because "discovery" is a term which refers to an epistemic achievement: if one succeeds in discovering something then, no doubt, this something exists. (Arabatzis, 2006, p. 217)

For instance, as Arabatzis (2006, p. 217) points out, these days one would hardly claim that phlogiston was discovered in the 18th century even though at the time many scholars regarded it to be a substantial discovery in chemistry.

Arabatzis argues that there is a distinction to be made between *discovery* and *generation* or *construction*. The difference, he argues, between these terms is their relationship to *truth*. While generation and construction might lead to truth, discovery already "implies truth" (2006, p. 218), which, Arabatzis concludes, makes discovery an "extended process, which involves both generation and justification" (Arabatzis, 2006, p. 226).

However, this also leads many authors of the above-mentioned volume to the conclusion that the DJ distinction can still be upheld in the sense that there is a difference between "original historical model of hypothesis generation and the 'final' form of justification" (Arabatzis, 2006, p. 218). Steinle, specifically, notes that though in practice scientists are very well "aware of the historical nature of their enterprise" (2006, p. 189), retroactively, scientists generally remove, as much as possible, any kind of reference to the process once it comes to communicating research achievements to others scientists and the wider public, such as in journal articles or textbooks – a process also noted by Reichenbach himself (1938, p. 6). This "process of decontextualizing" is also an attempt to portray the obtained knowledge as secure as possible, "i.e., stripped from the specific time, place, and process by which it has be generated" (Steinle, 2006, p. 189).[102] In this sense, Steinle argues, is the DJ distinction part of science itself.[103]

When it comes to climate science which substantially relies on highly complex computer simulations, the notion that a theory or hypothesis can be validated (at some point) independently from understanding the circumstances under which it was generated is not so easy to sustain. As shown at several stages throughout this chapter, the behaviour of a model is often intricately dependent upon its history. Understanding the history of the models is, therefore, constitutive to knowing to what extent a 'good' model output is actually rooted in a good representation of the relevant process in the model or just an artefact of some unknown interference between different elements of the model or some effect from tuning. However, this is easier said than done. As

[102] Steinle (2006, pp. 190–193), however, also remarks that the wider conceptual framework of a theory cannot be seen as independent of its historical origin. While these overall concepts (an example from history of science is the concept of absolute space in physics) are unconsciously followed in daily scientific practice, they are also broadly contingent upon the (social) context in which they were established. Nevertheless, depending on the point of view, Steinle argues, a theory can be considered to be justified separately from the context of its generation (that is, when the contextual framework is not taken into account).

[103] Steinle sees a form of Hoyningen-Huene's lean version in this insofar as scientists are posing two different questions: "how did a certain insight (a theory, law, fact, …) come about? And why should we believe it, what are the reasons for support?" (Steinle, 2006, p. 188) once they turn from the research process to communicating and teaching their findings.

we have seen in this chapter, global climate models have some degree of opacity to them, which entails that conventional methods of justification such as a 'simple' fit to observations or more complex ones such as V&V only have limited applicability.

On the other hand, Winsberg (2018, p. 160, 2003) argues, with reference to Hacking (1983), that the techniques applied in complex computer simulations,[104] as used in climate science, have "a life of their own"; meaning that they are justified not just by theory alone but also because they are understood to be well-established procedures:

> Whenever these techniques and assumptions are employed successfully, that is, whenever they produce results that fit well into the web of our previously accepted data, our observations, the results of our paper and pencil analyses, and our physical intuitions, whenever they make successful predictions or produce engineering accomplishments, their credibility as reliable techniques or reasonable assumptions grows.
>
> That is, the next time simulationists build a model, the credibility of that model comes not only from the credentials supplied to it by the governing theory, but also from the antecedently established credentials of the model building techniques developed over an extended tradition of employment. (Winsberg, 2003, p. 122)

Thus, in Hacking's (1983) words, these techniques are "self-vindicating" or, as Winsberg puts it, "they carry their own credentials" (2003, p. 121). That is, similar to Hacking's claim about experiments, the credibility of these techniques lies, according to Winsberg – at least to some parts – in their historical successful application.

In practice, a certain knowledge about the history of the model can also be beneficial insofar as it can reduce the opacity computer simulations of this

[104] Winsberg specifies that by "techniques" he is "referring to the whole host of activities, practices, and assumptions that go into carrying out a simulation. This includes assumptions about what parameters to include or neglect, rules of thumb about how to overcome computational difficulties—what model assumptions to use, what differencing scheme to employ, what symmetries to exploit—graphical techniques for visualizing data, and techniques for comparing and calibrating simulation results to known experimental and observational data" (Winsberg, 2003, pp. 121–122).

type bring along.[105] If within a community of modellers the knowledge about some past modelling decisions gets lost over time, it can lead to problems in the future when some new adjustment results in a seemingly unexplainable change in the model's behaviour. There might be a situation, where a haphazardly constructed model implementation due to limited computing power is introduced to the model at a certain moment in time but forgotten after some generation of scientists as it has not interfered negatively with any other model adjustments in the meantime. Lack of knowledge of the history of the model can become a significant obstacle for scientists when a new improvement to a model component all of a sudden does interfere with the previous imperfect and forgotten adjustment and leads to an unexpected bad model performance (Lenhard, 2018, pp. 839–840 describes an example of such a case). If one sees models as instruments applied to gain knowledge about the climate system, as many climate scientists do (Chen et al., 2021, p. 215), then knowledge about the history of the model can strengthening the confidence in the instrument.

All of this makes it questionable if the notion that the procedure of justification can be fully uncoupled from the historical circumstance is still viable. It also shows how the expertise and experience that scientists have acquired throughout working with these models and the techniques employed to develop them are key to assessing the models and resulting hypotheses.

3.4.3 Scientific objectivity

The third concept that was pointed out in Chapter 2.3 was *scientific objectivity*. The ideal of value-free science is often closely connected with the concept of 'objective' science. The problems of this interpretation of *objectivity* were examined in Chapter 3.1. It was shown that a definition of *scientific objectivity* as complete value-freeness of science cannot be maintained in actual scientific practice. An alternative definition of scientific objectivity that is derived from a plurality of perspectives has been shown to have more chances of being successful (Leuschner, 2012a; Longino, 1990). Another related application of the terms objective and subjective we have seen in this chapter refers to the fact that no

[105] That disclosing the specific circumstances and methods of model development can improve model evaluations has also been raised by scientists themselves. For instance, Mauritsen et al. (2012) advocate for making the tuning process more explicit in inner-scientific discussions and communications.

perfect model exists and some subjective decisions have to be made concerning which and how specific processes are included (Tebaldi and Knutti, 2007). A further use of the term objective we came across in this chapter concerns a simplified description of the relationship between observations and theories, where observations supposedly provide 'objective', irrefutable evidence that a theory (or model) is right or wrong.

Here however, I would like to take another look at a somewhat more specific conception of *scientific objectivity*, which is more common in climate science. Climate scientists particularly evoke the term *objectivity* when describing mathematical or automatic procedures. *Subjective* approaches are marked by being 'done by hand' and relying on expert judgement. This distinction of objectivity and subjectivity for instance is often used in descriptions of the tuning process (see also Chapter 4.2.2). By objective method "one means that a well-founded mathematical or statistical framework is used to perform the model tuning, for instance, by defining and minimizing a cost function or by introducing a Bayesian formulation of the calibration problem" (Hourdin et al., 2017, p. 594). Contrary to this, the more common approach to tuning is described by climate scientists as "subjective" and is more directly guided by the expert judgement of the scientists. However, as scientists note, application of the afore mentioned objective procedure still has subjective components to the extent that "[a]ny such objective tuning algorithm requires a subjective choice of a cost function and this involves weighting trade-offs against one another" (Mauritsen et al., 2012, p. 16).

Although the scientists note that subjective judgements (at this moment) are unavoidable, it is also discernible that *objectivity* for the scientist has a positive connotation and is to be preferred to subjective methods. The hope put into objective procedures is to find a way to exclude personal influence on research. Subjective methods imply that the scientist has to make some kind of judgement. To make these judgements scientists have to rely on their expertise and experience. But the worry is that this inevitable reliance on this expertise-led decision-making will, considering the general complexity of climate modelling further "an unfortunate reduction in transparency" (Schmidt et al., 2017, p. 3208). Appealing to objective methods is linked to the wish to bring transparency, traceability and reproducibility into these processes.

3.4.4 Conclusion: what now?

What these three concepts have in common that they allude to the relevance of another element of doing climate science that, again, is not new to science but has been gaining significance in the context of the epistemic challenges rooted in the increasing complexity in science: the *experience* of scientists in their field in general and especially in dealing with the particular instruments and methods (for example, the specific simulations they use).

This experience is what Lenhard (2018) describes as the "feeling" scientists have for the models, which can only be acquired in practice, in working in their area of expertise and in working with the models.

In a similar vein, one might interpret the emphasis philosophers and scientists place on understanding the history of the models and modelling practices. Computer simulations of this kind cannot be epistemically grounded fully in an ahistorical way. On the one hand, as Winsberg has pointed out, the credibility of the model is based in part on the techniques and practices of model construction through the tradition of their application. On the other hand, knowledge about the history of the models themselves can be crucial to circumventing, at least to some extent, the opacity the complexity of the models causes. Both cases highlight the relevance of the experience that the scientists have with handling the models in practice to increasing the credibility of models and modelling techniques.

Climate scientists themselves have also observed the significance of experience to their work (Tebaldi and Knutti, 2007). However, as we have seen in this chapter, in this context the notion of expertise, experience or expert judgement is also often associated with subjectivity. Though there is often a negative connotation to subjective decision-making, climate scientists also note its unavoidability.

But, when experience has such a significant relevance, then three questions have to be asked:

1. about the nature of this experience
2. how it is acquired
3. how it is justified.

These questions will be discussed in the next chapter. I will argue that a crucial element of this *experience* is tacit knowledge. A fundamental part in gaining expertise in any subject is grounded in experience, in having practiced in the specific field in question and often learning from working in close proximity to others already having some expertise. It requires a knowledge that is more than what can be learnt from books and it is not easily put into words without any kind of 'showing' either. I will employ a broad interpretation of the term *tacit knowledge* that does not only apply to knowledge which cannot be expressed in principle but also the kind of knowledge that is not expressed in practice for whatever reasons. Tacit knowledge is not new to science, however, so far it has lived a mostly inconspicuous life in science. But, as I will argue in the next chapter, it gains greater significance in the context of highly complex systems where traditional approaches to knowledge acquisition are challenged.

The relevance of tacit knowledge also steers the focus onto the institutions that are at the centre of the scientific community because they are the place where this tacit knowledge is acquired, taught and communicated. *Institutions* has to be understood here in a broad sense, not just as specific organisations but as the extensive structures that make the scientific endeavour possible.

Winsberg (2018) has made a similar point by arguing for philosophy of climate science to turn its attention to the social structures of science. But he comes to this conclusion from a slightly different angle. Winsberg concludes, because of all of the epistemic problems and obstacles in climate science already discussed in this chapter,

> that philosophers do better to paint a picture in which we urge trust in the consensus of the scientific community, based on features of that community's social organization, then to try to provide a normative framework from which we can demonstrate the reliability (or its absence) of such-and-such modelling results. (Winsberg, 2018, p. 161)

By this Winsberg does not negate that it is, for instance, possible that "a simulation modeler could explain to his peers why it was legitimate and rational to use a certain approximation technique to solve a particular problem" (Goodwin, 2015, pp. 342–343; Winsberg, 2018, p. 161). But this kind of inner-scientific process of legitimising certain techniques, methods or hypotheses is always local and specifically context-dependent. Instead of trying to find schemes that would ground climate models normatively, philosophers, Winsberg argues, should rather focus on the social structures as "climate science is, in a thorough-going way, a socially organized kind of science, and […] many features of its epistemology need to be surveyed at the social level in order to be properly understood" (Winsberg, 2018, pp. 209–210).

In the following I will also argue for turning the attention to the social organisation of science. For one, I agree with Winsberg that the social structure of science – specifically one that is as scattered between so many different individual scientists and institutions, dealing with highly complex systems and resulting epistemic problem as climate science – is significant to comprehending and grounding our trust in it. But I will argue as well that the increasing significance of tacit knowledge, or what Lenhard calls a "feeling" for the model, will also give new epistemic significance to the social structures in science.