2. Modelling, Qualitative Models, and Model-Based Diagnostics

This chapter will focus on the topic of modelling. It plays the important role of proposing an understanding of modelling that, as I will argue in the next chapter, maps onto the previously established picture of psychiatric diagnostics. This in turn will establish my proposal, the *model-based account of psychiatric diagnostic reasoning*, as an answer to the Methodological Question.[1]

While a whole chapter on modelling may seem excessive at first, it is crucial. It is crucial to give space to development the framework for modelling that I intend to apply to psychiatric diagnostics, because the proposed understanding of modelling has to meet specific requirements, mentioned in the Introduction to this thesis, to provide a methodology of modelling that, if successfully applied to psychiatric diagnostics, provides a satisfying answer to the Methodological Question. To recap, it needs to provide a description of the method assumed to be at work in psychiatric diagnostics, it has to provide an understanding of the rationale for the method to operate the way it does, and it has to speak to us about why and how the conclusions of the method may be deemed justified. Only when an understanding of modelling that can address all these points is established will an attempt to map the proposed method of modelling onto psychiatric diagnostics be able to yield a qualified answer to the Methodological Question. To generate this fully developed account of modelling, an entire chapter is required. Let me next sketch how the chapter is set up.

I begin this chapter by presenting a description of the type of modelling that I take to be realised by psychiatric diagnostic reasoning – namely, qualitative diagnostic modelling. To this end, I first (2.1) provide a general a general account of modelling, distinguishing it from other kinds of theorising based on contemporary debates in philosophy of science. Next (2.2), I introduce a specific format of modelling, qualitative modelling, as well as (2.3) a certain application of modelling, diag-

1 I have already begun to think about a model-based account for psychiatric diagnostics in Kind (2023). As the reader familiar with my previous work will note, the understanding of the type of modelling I discuss changed and evolved since my earlier reflections on the topic though I still take d Godfrey-Smith's (2006) and Weisberg's (2007; 2012) work as a starting point.

nostic modelling. After providing this description of the relevant type of modelling, the remainder of the chapter focuses on three things. First (2.4), I analyse the inferential strategy used in diagnostic modelling to provide an epistemic understanding of the rationale behind diagnostic modelling (why this kind of modelling proceeds as it does), and spell it out in terms of what I call the constitutive-indicator strategy. Second (2.5), I discuss the types of inferences executed by following the constitutive-indicator strategy, which I argue to be *inferences to the best explanation*, *apophatic inferences*, and *inferences to unintelligibility*. Third and finally (2.6), I discuss to what extent these inferences occurring in diagnostic modelling may justify its conclusions. I conclude (2.7) with a brief summary of the chapter.

2.1 Modelling

My general understanding of modeling is a slightly modified version of Godfrey-Smith's (2006) and Weisberg's (2007; 2012) accounts, with the latter building upon the former. Their accounts were derived from case studies in evolution and population biology and informed by previous debates on modelling, mainly in the philosophy of physics and economics (e.g., Cartwright, 1983; Wimsatt, 1987; Giere, 1988; Morgan & Morrison, 1999). Currently, their view is not only highly plausible, it is also the most comprehensive and detailed account of modelling as an epistemic practice in the philosophy of science. Therefore, their account is a strong candidate for determining whether a certain epistemic practice, such as psychiatric diagnostics, should be classified as modelling.[2]

Godfrey-Smith's and Weisberg's main idea is that theorists developing and using models (i.e., modellers) follow a particular strategy of theorising to develop theoretical models of empirical systems. They call this strategy the *indirect strategy of representation* (from now on ISR). A theorist following this strategy engages in a three-step procedure. First, they set up a theoretical structure based on limited initial information about the target system and assign aspects of this structure to an element of the targeted real-world system. Second, the theorist investigates the properties of these theoretical structures to learn about its dynamics in order to predict its future states and outputs. Third, the theorist compares the findings of the structure's properties to the behaviour of the real-world system(s) that the theoretical structure was intended to target and judges whether the structure can be used to satisfy the theorist's epistemic interest in the system, for example by predicting changes or simulating entire processes taking place within it.

2 Another feature of this account making it attractive for dialectic reasons is that it is free of controversial commitments regarding the ontology of models and theories of model representation (Frigg and Hartmann, 2020).

If it turns out that the structure enables the theorist (and other competent users) to make this inference with sufficient precision to meet their needs, the structure is accepted for usage and considered to be a credible model.[3] If the structure does not meet the pragmatic criterion of being useful to the modellers' aims, it will be rejected and either another structure is set up or a modified version of the already tested structure receives a second round of analysis and comparison to achieve credibility. The individuals following these steps are automatically considered to be modellers.

To contrast modelling (i.e., following the ISR) with other forms of theorising, Godfrey-Smith and Weisberg introduce an approach to theorising that they call the *abstract direct strategy of representation*. While following the ISR procedure is modelling, following the abstract direct strategy of representation is supposed to result in data-driven theorising, which in their understanding is distinct from modelling. A theorist following the abstract direct strategy of representation proceeds as follows. They begin their pursuit of a theoretical structure that targets real-world systems by generating and collecting large amounts of available data about the system(s) of interest. This way they address it *directly* before they begin to theorise. Based on the large amounts of data they collected, they try to determine which properties of a system appear to be essential to account for other properties of the system and set up a theoretical structure based on this judgement. In this process they abstract the rest of the data not needed for this purpose. In the end, they arrive

3 Credibility is a matter of pragmatics rather than truth (in the sense of faithful/complete representation), and it is the central goodness criterion applied to evaluate models (e.g., Sargent, 2010; Truran, 2013). For a model to be credible it is neither *necessary* nor *sufficient* for it to be a faithful and complete representation. It is not even typical. It is not necessary because many if not all predictive models and simulation models are highly idealised, but they are nonetheless sufficiently predictive and/or simulatively accurate to be used in alignment with the modeller's interests and therefore credible. Faithful and complete representation is not sufficient because credibility is a pragmatic matter, and a faithful representation may under some circumstances not be the right tool to archive the aims of a modeller. If, for example, a modeller wants *one* model that makes a prediction (with some margin of error) about multiple similar but in many regards different systems, it seems possible that no faithful representation of any of these systems – and there can be *one* faithful representation of several different systems – could provide the modeller with a "model" credible for the task of making the desired predictions across this systems, while it may be that a model that contains idealisations is in fact well suited to the task. Note that the latter illustrates a *contingent* and not a *necessary* tradeoff between precision and credibility resulting from the interests of the model. I take this to be uncontroversial, whereas claiming a necessary tradeoff relationship between precision and utility would be a highly controversial claim (Odenbaugh, 2003; Orzack, 2005, 2012; Matthewson and Weisberg, 2009). Finally, representational fidelity or completeness is *atypical* for what we call models, since in speaking about models we do not consider models to be anywhere close to being faithful complete representations of their target, but at best *partially* true representations of their targets (e.g., Da Costa and French, 2003).

at a theoretical structure meant to represent the target system's properties and their relations faithfully.

The crucial difference between both ISR and the abstract direct strategy of representation is that while the representational features of both theoretical structures are indirect in the sense that they end up proposing a theoretical structure serving as a vehicle for reasoning about a targeted system, ISR is indirect in an additional sense. It takes a deliberate extra step of setting up a theoretical structure to stand in for the system of interest and investigating this structure to learn from it, and *only then* relates this setup structure to the target system in order to evaluate one against the other. The abstract direct strategy, on the other hand, begins by directly collecting vast amounts of data about the intended target system, investigates the data, abstracts the materials that are unhelpful for arriving at a representation of the features of the target system that the theorist is interested in, and then sets up the theoretical structure on the basis of the retained data.

Now that the difference ISR and the abstract direct strategy is clear, let me add that the difference appears gradual rather than categorical. It is rare that a theorist sets up a theoretical structure meant to target a real-world system without *any* knowledge of the system. At a minimum, the theorist must have knowledge of the existence of the system, knowledge that gives reason to be interested in it, and some assumptions about it that leads them to set up one or another structure to target it. On the other hand, even those theorists who engage in some form of data-driven theorising about systems have a disciplinary background that provides specific structures typically used for theorising. Likely, those structures will be applied to analyse data that do not in themselves tell us how to order and analyse the data or make inferences about the real-world system based on this data.

Equipped with an overview of the principal strategy followed by modellers, let us go into more detail regarding each step of the modelling process: how we construct models, how we analyse them, and what we can learn from them about reality.

2.1.1 The Construal

The first step in this process is constructing the model system that is meant to target a real-world system. This step might be guided or inspired by existing theoretical sources providing full or partial structures for the model, by the limited knowledge about the target system, and by presumptions about principles that may govern aspects of the system or require the modeller to draw on previous experiences from (un)successful attempts to model similar processes or systems. Bringing all these sources of inspiration in play has been called the "*art*" or "*know-how*" aspect of modelling for which no manual exists (Morrison, 1999; Godfrey-Smith, 2006). It requires experience and expertise in modelling. The step of construal itself is characterised by four aspects. These aspects are not considered to be steps that have to be carried

out in chronological order; they are interdependent and must be considered simultaneously by the modeller. The four aspects are: finding a structure for the model, assigning the model to a target, determining the scope of the model, and setting up its fidelity criteria.

Structure: To construct a model, a theorist must select a theoretical structure via which they present the model. Such a theoretical structure may be quantitative or qualitative. It could consist of graphical representations, mathematical formulae, or interrelated propositions expressed in written text. Because one and the same intended underlying model could be expressed in different theoretical structures (i.e., a box-and-arrow diagram versus a formula versus a verbal description), while we assume each of these expressions to represent the same model, the chosen structure put forward will not be *the model*; it will be only one possible *description* of the model.[4]

In choosing a description, modellers will often be attempting to choose the one allowing them to capture the intended model's elements and relationships as precisely as possible. Moreover, the modeller will consider how reasonably a chosen structure will be able to be exploited in targeting the specific aspects of real-world system the model is intended to target (see "Assignment" and "Scope" below); what inferences to make about the system the structure is meant to enable (see "Fidelity" below); and what resources the modeller has available in order, later on, to compare a given model to the real-world system (see "Model/World Comparison" below). To come up with a model structure meeting all these requirements, the modeller can call on various sources.

One source employed might be the modeller's intuition, fuelled by methodological training and ideas about the target system. Another common source of inspiration for model structures are existing model structures, from the same or other branches of science, that have been used to model similar phenomena.[5]

4 Recognising the difference between the use of model descriptions rather than the actual model is not only plausible; it is also helpful for avoiding metaphysical questions. It is plausible, because not making this distinction would have implausible consequences. We would have to say, for example, that graphical illustrations of models that are also mathematically presented in scientific papers are not two descriptions of the same model but two different models. That is surely not what is intended by the scientists. It sounds more plausible that both the maths and the graphics describe the same model. Moreover, by admitting that all we are really dealing with in the process of modelling are different forms of model descriptions, we can avoid deep metaphysical discussions about models, such as whether in the end all theoretical models are mathematical and thus whether it is correct to speak of models that are not mathematical. Such problems can be avoided by the plausible assumption that what we encounter, modify, and handle in modelling are just descriptions, whatever the deep metaphysical truth about models might be.

5 The first kind of model reuse is called cross-contextual modelling (Knuuttila and Loettgers, 2016). A famous example are the Lotka-Volterra equations, first proposed to model autocat-

Finally, model structure may be based not on existing models but on assumptions articulated in theories that seem promising for addressing the relevant aspects of the target system. However, while reusing a model structure is often fairly straightforward, using a theory to come up with a model structure can be tricky. Sometimes theories make claims that are too abstract and do not in themselves provide specific structures for their application. Rather, an applicable structure has to be engineered based on the theory's principles.[6] Sometimes neither *one* theory nor *one* model provides the modeller with a coherent structure that appears a plausible candidate to map onto the regularities of interest in the system. In such cases, the modeller must draw on multiple models and theories providing partial structures or a basis for such structures, and this collection is then 'pieced together.'.[7]

Assignment: The assignment process accompanying the choice of the model structure specifies the model's target systems(s). In other words, it determines what the model is a model of, as well as which parts of the model structure target which aspects of the target system. Depending on how the model structure is specified, this assignment process might come more or less naturally. The assignment is more noticeable when the specification of the model's structure itself contains obvious hints – for example, if this model description contains symbols referring to aspects of the real-world system they are intended to be assigned to. For example, think of the small pictures of animals in a typical presentation of the "tree of life" model of evolutionary history, or the boxes of a model from a neuroscience textbook with brain areas names in them. Things are less straightforward, however, in models that are expressed in purely quantitative terms. These models need a more explicit articulation of their intended assignment. This problem is often solved by modellers through

---

alytic chemical reactions (Lotka, 1910) and later applied to model predator–prey interactions (Lotka, 1925; Volterra, 1926) and economic fluctuations (Goodwin, 1967). The second kind of model reuse has been described in terms of hub models (Godfrey-Smith, 2009). For example, suppose you have a detailed understanding of how one particular trait became selected in an evolutionary process. In that case, you might attempt to apply the same structure to model another trait's selection.

6 For an example, think of social network theory. This theory does not provide you with a specific network model to apply, but instead tells you how to set up a structure that complies with the theory. Or, to take a classic example, classical mechanics does not provide you with a model of a pendulum, but gives you the tools to develop a model structure for a real-world pendulum by providing a framework for taking different real-world factors (e.g., friction) into account in attempts to develop a model (Giere, 1999).

7 A nice discussion of several of these "puzzling examples" can be found in Boumans (1999) in the context of modelling in economics, where modellers attempt to integrate different sources (e.g., economic models, phenomenological laws, and assumed economic "laws") into business-cycle models.

conventions as to how elements of model descriptions are meant to map onto aspects of the model system. In physics, for example, there are established lists of constants. Everyone familiar with classical mechanics looking at a mechanical process model knows that *F* is assigned to force. In medical and social sciences, it is common knowledge that *n* stands for the number of subjects in a sample.

Scope: The third component of the model's construal is selecting its scope to determine which aspects of the phenomena will be targeted. This interest-guided process is the other side of the assignment process. The assignment establishes the intended mapping of the model onto the real-world system by telling us what real-world system the model targets and what parts of the model are meant to relate to selected aspects of the target system. The model's chosen scope determines the target systems' aspects to be targeted by the model and which aspects of the target system are left aside. The scope is usually determined by the modeller's interest and the presumptions about the aspects of the real-world system that will be relevant to achieve a credible model. For example, suppose you know that in ecological modelling of population growth, considering the amount of prey is an essential predictor in many models. In that case, as a modeller you might decide to take it into account for your own model of population growth.

Fidelity: The final aspect of the construal is the stipulation of fidelity criteria. They define adequacy conditions for models. Fidelity can be divided into two types. First, a model's *dynamic fidelity*, determining how similar the output of a model (its prediction) has to be to the real-world system's output to be considered credible. This criterion may take a numerical value and ordinal positions or take the form of quality space, and in the first case it is often provided with an error tolerance value regarding its outputs. The second kind of fidelity is *representational fidelity*, determining to what extent the model structure allows for simulations of changes that also occur in the aspect of the real-world system to which its elements are assigned. This is considered important if the purpose of the model is not only to predict but also to track causal pathways in the modelled system, usually leading to higher credibility requirements when it comes to representational fidelity.

2.1.2 Analysis of the Model

The second stage of the modelling process is the investigation of the model system itself. In this step, modellers familiarise themselves with the model structure they have developed and with its dynamics and predictions. In other words, they learn what elements are present in the model and what changes to which elements of the system lead to what changes in other elements.

This analytical step is logically autonomous from the model's application to a target system. For now, the modeller is concerned only with discovering regularities in the model structure. As long as the model passes the later step of model/world comparison, these regularities may suggest interesting options for using the model for epistemic and practical purposes in application to the targeted real-world system. Things that might be discovered in such analysis that may be interesting later include, for example, surprising relationships between elements in the model discovered in a simulation that might be useful in predicting changes in the targeted real-world system, or in guiding strategies to develop interventions to control changes taking place in this system. Moreover, successful models may even give rise to principles that might not only be true in the one specifically modelled system but might also turn out to be generalisable to other systems. As Weisberg puts it: "Even where the model is inspired by a real-world system, what the theorist finds out about it is distinct and usually more general from the system which inspired it" (2007, p. 19). Examples of such principles developed from a single narrow, targeted model include the "Volterra Principle", in which general pesticides (i.e., an intervention that kills both predator and prey) increase the relative proportion of the prey (Roughgarden, 1979), and "Dunbar's Number" (Dunbar, 1992; Hernando et al., 2010), stating that human groups larger than 150 will not be sustainable based on personal relationships. A feature of model analysis is therefore not only the promise of discovering regularities that might hold in the specific system it will be evaluated against, but also the potential discovery of more general principles applying to other relevantly similar systems.

2.1.3 Model/World Comparison

In the last step of the modelling process, the results of the analysis are compared to the real-world system to see how well the model's predictions and simulations map onto the aspects of the system targeted by the model. In such a comparison, the modeller compares the model not directly to the targeted system but instead to a description of this system. This description needs to represent the system's relevant features in a format that matches the format of the model structures, so that a comparison is possible.

The need to craft a theoretically adequate description of a system in order to bring it in touch with models has been helpfully identified in philosophy of science by Nancy Cartwright (e.g., 1989, p. 133). She introduced the difference between a *prepared* and an *unprepared* description. The unprepared description "contains any information we think relevant in whatever form we have it available […]. We write down whatever information we have" (ibid.). However, such a description does not usually allow a model to be assessed against it. For this we need a prepared description instead. In such a description, "we prepare the phenomenon in a way that

will bring it into the theory" (ibid.) – or in our case, the model. If, for example, we have created a mathematical model of the changes of temperature in the course of a chemical reaction, the description of the system (the reacting compound) will have to provide a representation of the target system's modelled aspects in a numerical form that makes its states and changes relatable to the proposed mathematical structure of the model. This means that if the temperature of the compound is modelled in terms of degree Celsius, the prepared description of the system's heat states will also have to be given Celsius, to allow for assessment of the model.

If such a prepared description of the systems is provided, the theoretical model is compared to it, and the modeller decides to accept or reject the model in light of the results of the comparison. The model will be accepted if the model matches the behaviour of the system well enough to meet the modeller's interest in predicting or simulating the system, and if it can therefore considered to be a credible model of the system. The "good enough" status is determined on the basis of the previously stipulated fidelity criteria. If the model mismatches the system – that is, the comparison reveals that the model does not meet the initially formulated fidelity criteria – the model will be rejected. However, if the modeller does not intend to end their efforts at this point, such a mismatch can itself be used to modify the model structure in an attempt to come up with a model that better fits the targeted system. If this route is taken, it requires another round of analysis and comparison with the model's target to establish whether the resulting model might be more acceptable. If not, more fitting work or the invention of a totally different model structure might be the path the modeller has to take.

So far, I have presented the basic outline of the modelling that I will apply to psychiatric diagnostics. In the next section I will flesh out a variation of one aspect of the modelling procedure – namely, the potential choice of a qualitative over a quantitative model structure, and how such a choice influences the modelling process and its results. Understanding these consequences will be important for our purposes here, because I will claim that models used in psychiatric diagnostics qualify as qualitative models.

2.2 Qualitative Models and Qualitative Modelling

Modelling as presented above can take place in the form of either quantitative or qualitative modelling. In this section I will explore qualitative models and modelling. This step will be important for my project, since I will propose that psychiatric diagnostics, if it is understood as a modelling process, should be considered to involve qualitative modelling. To argue this point, a good understanding of the features of qualitative modelling is needed.

What makes modelling qualitative or quantitative is the nature of the employed structure. In this context, the *structure* of the model refers to the elements of which a model consists and the relationships between them. These elements and the relations between them can vary in nature. Most examples of models discussed so far, with some exceptions such as the "tree of life", were cases of scientific modelling in which the model structures are quantitative; their structure consisted of elements and relationships that are mathematically continuous variables. Qualitative modelling, on the other hand, is a form of theoretical modelling in which aspects of real-world systems are represented in a discrete and symbolical manner, no matter whether the real-world system is considered to be continuous or not. The chosen values introduced for the purpose of being assigned to states of the modelled system will in qualitative models usually be of limited number – e.g., "present", "absent", "neutral" – rather than being continuous variables with a potentially infinite number of values. Likewise, the relationships take the form of qualitative values such as "increases", "decreases", or "irrelevant" rather than indicating a quantitative measure for the influence of one variable on another. As a result of such discretisation of values in the model structure, each value can be understood symbolically, as making a reference to a discrete state or condition in the system. The presentation of the resulting qualitative model can take various forms. Such models can be presented in a drawing accompanied by guidelines on how to interpret it, as the "tree of life" model may be; they can be presented via the box-and-arrow diagrams we find in textbooks; or they can be represented in the form of conditionals in propositional logic. Qualitative models can also be presented as a set of interrelated propositions that are expressed by means of natural language, specifying the elements and relationships in the model. This last format for construal of a model has been called a *propositional model* (Thomson-Jones, 2012).

A typical and philosophically interesting feature of qualitative model if compared to quantitative models is their higher degree of idealisation. This higher degree of idealisation takes place in two forms that are terminologically differentiated in philosophy of science as *Aristotelian idealisation* (Batterman, 2002) and *Galilean idealisation* (McMullin, 1985).

Aristotelian idealisation is introduced to a representational structure in the context of determining the scope of the model. In this kind of idealisation, decisions are made about which features of the real-world system are intended to be represented by the structure and which will be abstracted from. The more is intentionally left out, the higher the degree of Aristotelian idealisation. If, for example, I develop a model of the population dynamics in an ecosystem but knowingly ignore certain populations that I know to be present (e.g., smaller animals like insects), I engage in Aristotelian idealisation. While such forms of idealisation take place in any kind of modelling, it is typical for qualitative models to show a higher degree of it due to the more limited number of elements and relations that usually appear in them

compared to quantitative models. Of course, qualitative models *could* in principle have infinitely many elements and relations with infinitely many discrete qualitative states of parameter and relations so that the degree of Aristotelian idealisation could be decreased, but in practice this would undermine one of the main reasons why many choose this way of modelling – namely, the computational simplicity that grants cognitive tractability to its outcomes.

Galilean idealisation takes the form of deliberate distortions to the representational structure that targets aspects of the real-world system, to address them in a simplified way. For this, the elements and relationships in the structure undergo simplification that intentionally reduce the complexity of the targeted features. An example of such simplification would be developing a model showing how employment and education influence the likelihood of being in a long-term relationship, in which the variables and their influences represent known real-world complexities in a simplified manner. This may mean, for example, that education is differentiated only in terms of what degree one has, ignoring educational performance differences among people with the same degree, or that the variety of existing employment situations is reduced to the opposition between employed and unemployed. Regarding the relations between these elements, simplification might take place when the model assumes that being employed leads to a higher likelihood of being in a long-term relationship but intentionally ignores further factors that would complicate modelling this likelihood relationship, such as temporal factors (e.g., how long someone is employed) that the effect might depend on. While this type of idealisation also occurs in quantitative as well as qualitative models, qualitative models usually introduce a higher degree of Galilean idealisation than quantitative models. The usually limited number of elements and discrete values favours lumping together real-world phenomena that are in principle separable into fewer variables to ensure computational tractability, and relationships modelled as discrete qualitative states impair the capacity to address more complex relationships amongst variables. Given that both types of idealisation, Aristotelian as well as Galilean, are typically highly present in qualitative models, they are typically not the first choice of modellers interested in maximising the representational fidelity of their models. However, in a context where representational fidelity is not a central requirement for the model's use, qualitative modelling can have beneficial applications.

A central benefit of qualitative models is that they are cognitively more traceable and, *vice versa*, that they can provide a framework for a cognitively realistic understanding of expert reasoning.[8] Indeed, qualitative models not only *can* do this

8 However, it worth adding that qualitative models might also be chosen if suitable quantitative data about the target system that a higher-fidelity qualitative model would require are not available, or when the system itself is so complex that its computational intractability

but they do: one of their primary roles in research and practice is to understand and support the expert reasoning. Research on expert reasoning has shown that experts tend to think about the features of systems they interact with in qualitative terms. When thinking about quantities, motion, space, time, causation, or frequency, practical experts often do so without using abstract mathematical methodological frameworks, instead sticking to the qualitative categories of folk reasoning (Forbes, 2008).[9] This idea has been most prominently developed in the research on "qualitative physics", a branch of cognitive science investigating the use of qualitative models by engineers and other technical experts thinking about artifacts and their functions (see, e.g., Bobrow, 1984; Falkenhainer and Forbus, 1991; Weld and De Kleer, 2013). Similar research can be found in attempts to understand expert reasoning in economics and engineering as a form of qualitative modelling (Farley, 1987).

Considering what qualitative modelling has contributed to our understanding of expert reasoning in other fields, it also *prima facie* appears to be a promising candidate for understanding how psychiatrists approach their diagnostic task. As in the case of other experts who think about most of their work in qualitative terms, we can plausibly expect that the same is true for psychiatrists, since psychiatrists do not *calculate* their diagnosis, instead thinking about diagnostic matters mainly in qualitative terms. This becomes clear whenever one listens to diagnostic discussions at case conferences or in a clinical setting, from diagnostic discussion in training literature (e.g., Wright, Dave, and Dogra, 2017), and also from research on clinical reasoning that uses think-aloud protocols to show that clinicians think in qualitative terms about their cases (e.g., Audétat et al., 2012).[10] More about this will be said in the next chapter.

To bring my discussion of qualitative modelling to an end, let me ensure that we move on with a clear idea of how this kind of modelling works by providing an ex-

---

could render qualitative models better suited to predict or simulate aspects of the system that the modeller is interested in. Some recent examples of this later case can be found in areas of science dealing with immense complexity – for example, in attempts to model marine ecosystems, where highly idealised but tractable working models find applications (e.g., Reum et al., 2015).

9 Note here that this understanding of qualitative theorising differs from the understanding of Weisberg's (2004) discussion of qualitative theorising in chemistry. While qualitative models as discussed here share many features that he discusses too (e.g., a high degree of idealisation and a typically restricted number of variables), qualitative models as discussed here are not numerical, whereas Weisberg explicitly states that in his understanding, the difference between "qualitative and quantitative models is not about the use of numbers; both types of models can be numerical" (Weisberg, 2004, p. 1071).

10 For more on the validity of the use of this method, see Durning et al. (2013). Together, these forms of evidence seem to me sufficient to support the claim that if diagnostic reasoning in psychiatry is a form of modelling, it should be expected to be some sort of *qualitative* rather than *quantitative* modelling.

ample, adopted from Forbes (2008). Think about the process of heating and cooling water in a vessel on a machine that may increase and decrease the temperature of the water – that is, a freezer-heater. Assume that this machine has three positive and three negative levels (freezing, cooling, chilling, warming, heating up, boiling), and you are interested in the model of the conditions for water to take one of three qualitatively described states called "solid", "liquid", and "vaporising/boiling". These states are considered to appear on an ordinal scale such that any change between "solid" and "boiling" must pass the state "liquid". To develop a qualitative model for this real-world system, you may set up a structure with two variables, one assigned to the freezer-heater (saying what its current setting is) and one to the water in the vessel (saying in which state the water is). Then you assign a number of potential the qualitative values of the water and the heater-cooler setting to the elements of the structure. Which would make six potential values for the element mapping on the settings and three for the element mapping on the water. Then a relational structure can be set up, making claims about which qualitative value of the parameter assigned to the freezer/heater may lead to an influence on the parameter assigned to the water in the vessel. Finally, this model and its implications might be compared to the real world by playing around with the freezer-heater settings. Imagine that after some revisions that allow our model to match the real-world phenomenon, our model tells us that if the freezer/heater is on "freeze", the water goes down the ordinal scale stepwise until it is "solid". In contrast, the water goes up the ordinal scale until it is "boiling" if the machine is on "cook". In all other states, and the water will move towards (or stay on) the ordinal scale value "liquid". Suppose that these predictions are what can be performed by the model. In this case, we can make an accurate prediction following our initial interest in the relationship of the machine settings and the state of the water. And suppose that the water's transitions between the qualitative states assigned to them are also predictable in the model. In this case, we will also be willing to assign representational fidelity to it. The qualitative model meets the purpose of the modelling process without the need to set up a quantitative model of the system.[11]

11 What has been said so far should make clear the principal idea behind qualitative modelling, rather than its boundaries. Qualitative models *can* be more complex, and AI researchers and mathematicians have worked out frameworks to give technically more rigorous representations of qualitative modelling through qualitative algebra (Forbes, 2008). I do not intend to introduce and discuss any specific formal framework to talk about qualitative models; the educated intuitive understanding of qualitative modelling that should result from the previous presentation will suffice. What follows is therefore only gesturing in the direction of relevant work in AI and mathematics. For concrete examples of proposals for formalised qualitative models of complex systems, one may look at the examples of two-valued models, employed to diagnose dysfunctions in aircraft engines (Abbott et al., 1987) or photocopiers (Bell et al., 1994). A framework for employing three-valued formalisations based on a positive (+), nega-

Now that the outlines of qualitative modelling and its specificities have been provided, let us turn to the next specification of modelling that is relevant to its application to psychiatric reasoning: the use of models for diagnostic purposes, in the form of diagnostic modelling. In the next section, I discuss this specific use of modelling, and in the next chapter I will argue that this is the type of modelling that is also realised by psychiatric diagnostics.

2.3 Model-Based Diagnostics – Normative and Error Models

Modelling as understood in ISR varies not only in terms of the formats of models used (qualitative versus quantitative), but also in terms of the epistemic aims pursued through its application. Modellers may have specific interests in exploiting the regularities of the model system, evaluating the (dis)similarities between model structures and real-world systems, and assessing the outcomes of these evaluations. For instance, models that effectively capture variable changes over time can be utilised to predict specific occurrences in the modelled system, simulate its processes, or guide interventions to achieve certain changes in the system. As discussed with regard to the model analysis step, the ability of such models to match the modelled systems can be leveraged for a variety of purposes, highlighting the importance of careful model selection and analysis in ISR. Here I want to discuss another use of models: to identify and classify irregularities in the modelled system. The practice of setting up and using models for this purpose is called *diagnostic modelling*. If, as I argue, psychiatric *diagnostics* is to be understood as a modelling process, to consider it diagnostic modelling seems *prima facie* a plausible candidate. To further assess this plausibility, in the next chapter I will present this idea in more detail. In other words, in the remaining chapter, I will discuss what diagnostic modeling is and examine in the next chapter whether or not psychiatric diagnostics should be understood as a form of such diagnostic modelling.

The basic idea of diagnostic modelling was proposed by Reiter (1987). Diagnostic Modelling enables the decision as to whether an error of a certain type occurs in a real-world system via a comparison between this system's actual performance (in terms of outputs or internal processes given certain inputs at some point in time) with a presupposed model of the system, which I will call the *normative system*

tive, (-), or zero (0) value (on an ordinal scale) to model physical systems on different levels of complexity can be found in de Kleer and Brown (1984). Moreover, proposals have been made for the formalisation of monotonic relationships between model elements (e.g., if A goes up, B goes up) as well for compositional relationships (e.g., if A goes up, B goes up, *iff* C goes down) and how change over time can be considered in a time series of a qualitative model (Forbus, 1984).

model.[12] This normative system model is "a model which can be used to simulate the normal work of the system in the case of lack of any faults" (ibid., p. 440). Depending on how fleshed-out this normative model is, different aspects of the system at hand may be compared to it to find errors in the system. Given the normative system model and the actual system it targets, the modeller can then evaluate whether the real-world system shows the normal operations assumed by normative system model – that is, whether they can initiate a diagnostic process. In this diagnostic process, the modeller makes a "comparison of the observed system behavior and the one which can be predicted with the use of the knowledge about [the] system model" (ibid., p 440), which I call the normative system model. This way, deviations can be identified between the system's actual behaviour and how the system should behave under certain input conditions if it works without errors. Recognised deviations can then be diagnosed as errors.[13]

Based on this general idea of Reiter's, I want to propose some variations of model-based diagnostics that a modeller may engage in and that may lead them to make diagnostic statements of different levels of granularity. These kinds of diagnostic modelling will be relevant for understanding psychiatric diagnostics as modelling.

The first and simplest way in which model-based diagnostics takes place is what I will call *normative-model diagnostics*. In this kind of diagnostic modelling, the way to arrive at the conclusion that the system is in error is to look only at the inputs the system receives and the outputs it provides compared with what would be expected to be the output under the same condition in the normative model. In this method, all insights about the system error – including identification of errors and the entire

12 Reiter (1987) calls the model used for the comparison the "system model". I include the term *normative* to emphasise the model's function.

13 If you are familiar with recent developments in psychiatric research, you may wonder how the role of normative models as described here relates to recent uses of normative modelling in psychiatric research (e.g., Marquand et al., 2019). While both approaches make use of normative models, they do so for different purposes. The role of normative models in the account of diagnostic modelling proposed here is to enable the identification of errors based on deviation from an accepted normative model. Normative modelling in psychiatric research is "a class of emerging statistical techniques useful for *understanding* the heterogeneous biology underlying psychiatric disorders at the level of the individual participant" (ibid., p. 1415, my emphasis). Briefly and non-technically, this is meant to be done by establishing a "mapping between behavioral, demographic or clinical characteristics and a quantitative biological measure, providing estimates of centiles of variation across the population" (ibid., p. 1416). This mapping can then be used to try to identify biological variations found in individual deviations from the normative model. While normative models used in diagnostic modelling serve to identify deviations present in the system, the relevant use in psychiatric research is to support the discovery of biological variations co-occurring in individuals who are already assessed as deviating from the normal functioning determined in the normative model.

basis of a classification of them – are the system's outputs, which simply inform the modeller that the system itself must somehow deviate from its inner working and do not presuppose a detailed model about the system's inner constitution and normal functioning. A modeller may be satisfied by learning that the real-world system is *somehow* in error. They may be able to classify and identify system errors based solely on inputs and outputs, but they do not have to.

Diagnostic modelling has the potential to be more detailed;[14] the normative system model is not restricted to treating the system as a black box and only look at inputs and outputs. It may, in addition, address internal processes and components of the diagnosed system, yielding finer-grained diagnostic statements that do not identify a type of error based exclusively on unexpected input-output findings. Diagnostic modelling can serve to differentiate and identify errors based on what is occurring within the system when the erroneous output takes place. Such a closer look enables a more differentiated approach to identifying an error, based on a better understanding of how the system produces this error. This may even enable the modeller to differentiate between two types of error in a system that are indistinguishable in terms of an input/output relationship, but that differ regarding the system states presumably responsible for the erroneous output. In the latter case, this would allow the modeller to differentiate a *prima facie* singular error phenomenon into two errors of the system, which may lead to the use of two different diagnostic labels for them instead of one.[15] This higher level of detail in diagnostics is worth

14 At this point, one may ask where these normative models come from. They have to be established by previous system modelling efforts, under conditions that for the modeller community interested in the system were assumed to be normal working conditions.

15 Why do I say it may lead to more than one label? It might be the case that a modeller may set up a taxonomy with a one-to-one mapping between models and labels, but this is not necessarily the case. There might be reasons to establish a many-to-one mapping instead – from various errors to one and the same label. Reasons to prefer a many-to-one taxonomy may, for example, be pragmatic. Assume that the overall purpose of the labels is to guide interventions to repair the system, and among the considered interventions one type of intervention works just as well to return a system to normal functioning for multiple errors that lead to a similar erroneous output. In this case it would be a practical option – that is, one that allows us to succeed in our task considering the commitments we make in the attempt to fulfil it – to introduce only one diagnostic error label for the purpose of interventions. As Zacher (2002) pointed out with the example of psychiatric diagnostic labels, there are many practical commitments at work in coming up with a diagnostic classification system: "Deciding what counts as practical is complicated. With respect to categorizing psychiatric disorders, we should consider many things, including but not limited to available treatments; potential management strategies; the effects of labeling; maximization of true positives and true negatives in identification; establishing within-category homogeneity for creating groups in experimental research; uncovering etiologic scenarios (especially for spectrum disorders); mapping time courses; predicting prognosis; achieving coherence with basic science in genetics, physiology, and psychology; being both clinically informative and easy to use; and

pursuing not only for the sake of more precision in classification in itself, and the epistemic interest that might be related to this, but also because it promises pragmatic benefits relevant for modellers with interventionist interests: Finer-grained, more informative classification may aid in choosing interventions to target the relevant deviations of the diagnosed system and so restore normal working, or suggest how to modify the system to compensate if some parts of it are broken beyond repair. To achieve more detailed diagnostic modelling of this type, another kind of model is needed in addition to the normative system model; I call it the *error model*.

Error models are derived from theoretical background assumptions about *how* real-world systems deviate from the normative model when showing certain erroneous outputs. Just as in the case of the normative model of a system that has to be pre-established before diagnostic modelling can take place, error models are developed from the variety of sources described in 2.1 by the diagnostic modeller or their modelling community, providing them with a repertoire of models addressing not only the inputs and outputs of the system but also additional features that enable them to identify errors. The use of such error models presupposes, of course, that the normative model in contrast to which these errors are meant to be identified makes assumptions about how the system components or processes addressed by the error model should behave normally. If the normative model meets this requirement and error models for a certain type of error exist, three different kinds of diagnostics that uses error models may take place: *error-model diagnostics*, *differential diagnostics*, and *exclusion diagnostics*.

Error-model diagnostics is the most straightforward form of diagnostics employing error models. It can be used whenever a system provides an output that is suspected to be erroneous. Instead of just assessing the inputs and outputs (given the assumptions of the normative system model) to identify this error, the error model (assuming that only one error is believed to potentially apply given the output) is employed to assess the presence of this error by assessing the entire system in order to provide a diagnostic evaluation. This kind of diagnostics, while being more detailed in assessment, does not lead to different conclusions than normative-model diagnostics; it is merely a more precise way to come to the same conclusion

meeting psychometric standards such as reliability and validity. On some level, these are all practices. We do things with category members. We interact with them and based on that interaction we learn how to think about (or use) the category" (ibid., pp. 222–223). The same goes for diagnostics of systems in general, so that in the struggle to do justice to a long list of practical commitments like this, it seems plausible that pragmatic considerations may lead to a categorisation of ontologically different system errors under a common error label to do justice to its commitments. For my presentation of diagnostic modelling I will set this complex discussion aside, and just assume the modeller to be interested in a maximally differential classification system.

and apply a diagnostic label from the modeller's error taxonomy. It therefore provides no diagnostic advantage over the simpler normative model diagnostic process, which is likely to make it a rarely used approach for diagnostic modellers in practice.[16]

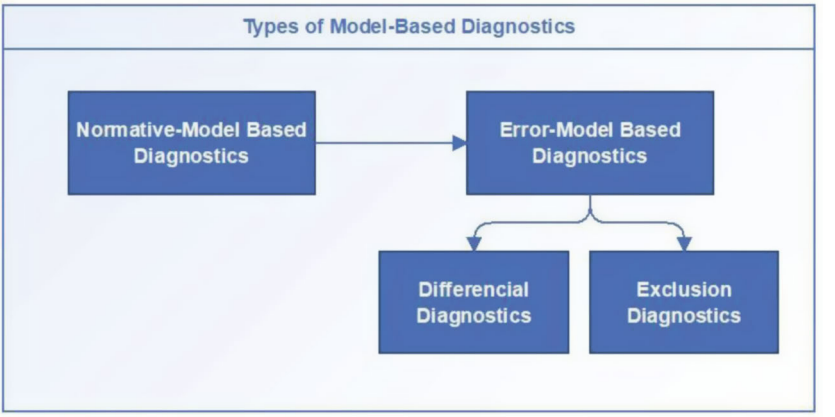
Differential diagnostics takes place if the system's output suggests to the modeller that there is more than one error model that might match the system that comes to produce a certain erroneous output. In this case, the modeller takes the error models potentially applying to the system given the recognised erroneous output and compares them to the prepared description of the erroneous system before making a comparative judgement as to which error model applies best (above a specified threshold of good fit) for this system and selecting this model. An error-diagnostic label from an error taxonomy that is associated with the chosen model is selected and applied to the system to diagnose its error.

Exclusion diagnostics takes place if there is more than one way in which a system may be bringing about a certain error, but we do not have an error model for each of these ways. In other words, the modeller works with a knowingly incomplete set of error models for a given erroneous output. This diagnostic process then starts out in the fashion of error mode-based diagnostics or differential diagnostics, but it contains the explicit possibility that none of the error models compared may match the diagnosed system. In such case –– i.e., if one or more error models were assessed without finding them to apply to the system –– the error label assigned to the system will lead to the classification of the error by exclusion diagnostics. An error with a label identified by this way of diagnosing is more information than if it had been classified based only on the inputs and outputs of the system, but also less information than if it had been identified by the successful application of an error model. The diagnostic process has provided information that leads to a partial negative identification of the error by excluding things that might be responsible for it.

In this section I have described the process of model-based diagnostics as the use of normative models and error models for the purpose of system error diagnostics. I have also mapped out different types of error-model diagnostics. We will go into more detail on these types of diagnostics in the next chapter, where I argue that psychiatric diagnostics understood as modelling implements this approach and therefore is a kind of diagnostic modelling.

16 If this seems hard to grasp, imagine a case from medicine. A patient presents with an illness for which we have a symptom-based method of assessment, but we could also do more detailed biological testing. However, based on our best understanding of clinical conditions and what the tests we have at our disposal can detect, the only thing that the test could detect is the same thing that we can identify based on the symptoms. In this case, it would be a waste of effort to engage in a more detailed biomedical examination.

Figure 8: The four types of model-based diagnostics discussed so far. Arrows indicate the presuppositional relationship between the different types of modelling. For differential or exclusion diagnostics to be possible, the presuppositions for error-model diagnostics must be met. For error-model diagnostics to be possible, the presuppositions for normative-model diagnostics must be met.



Up to this point, I have provided a stepwise description of modelling. I described modelling in general, I introduced qualitative modelling, and I presented a specific modelling procedure, namely diagnostic modelling. If we bring these elements together, the result is a qualitative modelling procedure for diagnostic purposes: *qualitative diagnostic modelling*. However, if I stopped here and moved directly to the next chapter to show that this kind of modelling maps onto psychiatric diagnostic reasoning, I would fail to answer the Methodological Question. All that this would achieve would be to provide a plausible reconstruction of the method used in psychiatric diagnostics. But this is only one of the three central aspects of the Methodological Question, as discussed in the Introduction.

What is also needed is an understanding of the rationale behind this method of modelling and an idea of how this method provides justification for its conclusions. Only if these two questions are addressed too can a demonstration that diagnostic modelling as presented here maps onto psychiatric diagnostics provide a full answer to all aspects of the Methodological Question. To ensure that my mapping attempt in the next chapter provides this full answer, I will therefore address these two aspects. Thus, when I demonstrate in the next chapter that psychiatric diagnostic reasoning should be understood as qualitative diagnostic modelling, how we should think about the theoretical rationale for following this modelling procedure and how it justifies the diagnostic conclusions will already have been developed. A successful mapping in the next chapter plus what follows in this chapter can, taken together, be considered to provide a full answer to the Methodological Question.

In the next section, I will begin this work by uncovering the inferential strategy behind diagnostic modelling, and therefore the rationale underlying this approach.

2.4 The Inferential Strategy of Model-Based Diagnostics

In this section I will explore the strategy of model-based diagnostics. When I refer to the inferential strategy of model-based diagnostics, I mean the common inferential process present in all instances of diagnostic modelling ensuring its truth-conduciveness. In other words, I mean the inferential process at work in diagnostic modelling that can reasonably be assumed to reliably led to the increase of true and the decrease of false beliefs resulting from it.[17] Understanding this strategy will, as mentioned at the end of the previous section, help us to understand why diagnostic modelling is set up the way it is, and thus to answer another aspect of the Methodological Question. The strategy at work in diagnostic modelling is what I will call it the *constitutive-indicator* strategy.

The label *constitutive-indicator strategy* derives from the idea that diagnostic modelling using normative and error models employs models that are constitutive in nature, and that differences between the system's actual behaviour and its expected operations are taken as indicators to apply certain diagnostic labels to the system. To spell out this idea and its implications, I will proceed as follows. First (2.4.1), I will discuss what it takes to be an indicator and why models in diagnostic modelling serve their epistemic purpose via being an indicator. Second (2.4.2), I will discuss what it means for a model to be a constitutive model, and how diagnostic models may be understood as constitutive models. Third (2.4.3), I will discuss the inferential patterns (inference to the best explanation, apophatic inference, inference to unintelligibility) that are realised in diagnostic modelling. Finally (2.4.4), in light of my previous discussion of the inferential strategy of model-based diagnostics, I

17 I focus on truth-conduciveness here, since it is usually considered the highest-ranking epistemic goal that should be supported by an epistemic practice. It is the "Epistemic Gold Standard", as Schurz (2011) puts it and as many other epistemologists also believe (e.g., Goldman, 1986, 1999; Bishop and Trout, 2005; Leplin, 2009, Ch. 2; Schurz, 2009). This position is not universal, however. Elgin (2017), for example, argued that the chief epistemic desideratum, especially in science, is *understanding* rather than truth, and that epistemic practices in science relying heavily on modelling are not aimed primarily at establishing truth in the first place. Going deeper into this discussion is beyond the scope of this project, but there have been several works that in my view convincingly refute Elgin's approach by providing alternatives to her understanding of the epistemic role of idealisation of science (Sullivan and Khalifa, 2019), unpacking the relationship between understanding and truth as epistemic desiderata, and defending truth-aptness as an epistemic priority including in science (Warenski, 2021).

will elaborate on the justificatory status of conclusions reached by a diagnostic modelling process.

2.4.1 Indication and Indicative Modelling

A widely accepted analysis of indication that I will adopt was provided by Dretske (1981). According to this analysis, an event of type *E* indicates a situation *S* to obtain if and only if the probability of *S* given that *E* occurs is 1, given some "channel conditions" under which these relationships are reliable. However, indication appears to be able to take place not only as an all-or-nothing affair; it can also come in degrees. An indicator may therefore be more or less reliable. To put it in terms of probability, some *E* may be an indicator for *S* with a probability anywhere between 0.5 and 1 [P(S|E)>0.5]. Different events, say *E1* and *E2*, might be better or worse indicators, depending on how reliably they indicate *S*. It seems necessary that some *E* must occur with a probability larger than 0.5 to be considered as an indicator at all. Otherwise, the "indicator" would not predict the absence or presence of a condition better than chance. You might as well flip a coin.

One feature of this understanding of indication is that the exact nature of the relationship, including the direction of the relationship, between *E* and *S* in which *E* is an indicator of *S* is undetermined. Concerning the possible nature of an indicator relationship, Dretske (1981) already noted that indicators may fulfil their role based on a causal relationship as well as a purely correlational relation.

The classic example to illustrate causal cases of indication would be the case that smoke indicates fire in certain channel conditions, since given these conditions, fire is usually the cause of smoke (fire smoke). However, this causal relationship does not enable indication only by looking at the later segment of the causal chain (smoke) to indicate the earlier one (fire); the relationship can also be used the other way around. Based on this relationship, we can also take fire as an indicator for smoke, treating the earlier segment (fire) as an indicator of the later one (smoke).[18]

18 The idea that indication may work up and down causal chains as well as via correlation is of course no philosophical achievement but has for many decades been a core topic in science interested in measurement. In psychology, for example, the terms *effect indicator* and *reflective indicator* have been around since Spearman (1904) first used factor analysis to measure general intelligence. He assumed that intelligence was the cause of the indicators he used to measure it, since changes in intelligence should lead to changes in the measured manifest indicators, but not the other way around. Today, most analysis using classical test theory, item response theory, or structural equation modelling shares this assumption. On the other hand, in psychology and social sciences we also find the term *causal indicator*, introduced by Blalock (1964), used to refer to manifest variables that can serve as indicators for the expression of a latent variable, which at the same time is theorised to be caused by this manifest variable. An example would be to consider the latent variable "life stress" to be reliably indicated by a manifest variable that would typically be interpreted to be causally

An example illustrating a correlational case would involve my dog, who rarely barks on any occasion other than when my doorbell rings, but almost always barks when that happens. Since my doorbell usually only rings when someone is standing in front of my door ringing it, the barking of my dog is a good indicator for someone standing in front of my door, but people standing in front of my door are not the direct cause of my dog's barking. People could stand there and not ring the doorbell, for example. But thanks to the channel condition that standing at my door usually goes together with it, the barking nonetheless correlates well with someone standing there. Regarding the variation in direction, in the case of a causal relationship the indicator may cause what it indicates, it may be caused by what it indicates, or it may just co-occur with it. Given this understanding of indication, let me finally say something about models.

A model might be thought of as an indicator of a state of affairs under two conditions. First, if a positive outcome of a model/world comparison using a specific model is usually reliably correlated with a given state of affairs in the system in a certain context. Under these conditions, the model's successful application can be an indicator for the relevant state of affairs. And second, if the inapplicability of a model in the context of a model/world comparison reliably correlates with a certain state of affairs, given certain background conditions. In this case, inapplicability can be an indicator of the relevant state of affairs. Let us now apply this basic idea of two forms of indication via models to see how it fits with the uses of normative and error models we encountered earlier.

As we have established, normative models and error models are the basic tools of diagnostic modelling. The attempt to apply a normative model to a system is used to indicate an error in a system, which is the case if the model is applicable to the system. Given what we said about indication, we can therefore now think of the inapplicability of a model as an indicator for the presence of a certain kind of error in the system in the context of normative-model diagnostics. Normative-model diagnostics therefore embodies one of the two ways in which models might be used as indicators: indicating a state of affairs *qua* inapplicability of a model. If we look at error models, and with them at the options of differential diagnostics and exclusion diagnostics, we find that both ways in which models might in principle be used as indicators play a role.

In the case of differential diagnostics, the modeller will consider which error model from their repertoire best applies to the system and take the one that is applicable as an indicator for a certain kind of error whose label is associated with the error model that was chosen. In this context, the applicability of a model is taken as

---

responsible for its presence, such as job loss, severe illness, or losing a loved one (Boolen and Davis 2009). For an insightful methodological debate on two conceptualisations of causally supported indication as well as covariance-based indication, see Bollen and Bauldry (2011).

an indicator for a state of affairs. For exclusion diagnostics, on the other hand, the same comparison procedure between the real-world system and the error models is pursued, but with the assumed possibility that none of the models may apply. The insight that no error model can be found to apply to the system so that no diagnostic label associated with one of the error models can be selected in these circumstances leads to an error label reserved for such an overall negative outcome being applied to the system. In these instances, it is the failure to apply one (or several models) that is taken to be an indicator to ascribe a diagnostic label to a certain state of affairs.

We have now discussed the indicator portion of the constitutive indicator strategy for diagnostic modelling, and thus established what indication is and how to think of models in general, and specifically normative and error models, as indicators. Now let us turn to the other portion of the strategy constitution. In the following discussion of this topic, I will show that the models used for indicational purposes by diagnostic modelling make constitutive assumptions about the phenomena that they are meant to indicate. As I will show, the models' constitutive nature is important because the resulting explanatory relationship between the models and phenomena ensures a reliable relationship that makes it plausible that they reliably indicate the targeted phenomena in the first place.

2.4.2 Constitution and Constitutive Modelling

The technical term constitution was already used in the last chapter with reference to Tyler Burge (2010, p. xv). Burge thinks of constitution in terms of constitution conditions – that is, the necessary and sufficient conditions for something to be what it is. In this context, a different understanding of constitution shall be considered. The understanding of constitution relevant here derives from debates about constitutive explanations rather than identity conditions. What is the difference? While constitution in Burge's sense concerns the conditions for something being what it is, addressing constitution in terms of a constitutive explanation aims to provide an explanation of a system's causal capacities, by giving an understanding of these capacities through reference to the system's parts and organisation (Ylikoski, 2013).[19] To understand a constitutive model, in other words a model that can be assumed to do explanatory work in terms of a constitutive explanation, a better un-

19 These two senses of "constitutive" differ regarding the *explananda* they can target, they involve different interpretations of what it is to spell out something's constituents, and they use different explanantia to provide answers to constitutive questions.
The potential difference regarding the explanatory target is quickly spotted. While constitutive explanations target causal capacities, Burge's account of constitution is not limited in scope in this way. Burge can ask what constitutes any phenomenon, including causal capacities but not limited to them. Burge could ask constitutive questions about the last fish in the ocean ("What is it to be the last fish in the ocean?") or Latin dance ("What is a Latin dance?")

derstanding of these constitutive explanations themselves is necessary. Therefore, I will proceed to flesh out the idea of constitutive explanation.[20]

---

that are excluded from being addressed by a constitutive explanation approach, since there are apparently no last fish in the ocean and being a Latin dance is not a causal capacity.

To see the extent to which the two approaches differ in their understanding of what it is to seek out something's constituents, we can imagine a constitutive question that both may address and see how it would be interpreted by them. For this purpose, think of the question "What makes the glass fragile?". Taking this question in Burge's sense would mean asking "*what does it mean to be fragile*?" and "*are sufficient conditions for fragility met by the glass*?". Interpreting the question in the sense of a constitutive explanation would mean asking "*why, in terms of its physical parts and their organisation, is the glass fragile*?". The first interpretation of the question is about spelling out what about the glass makes it true that it is fragile given the identity conditions of fragility, while the second question is about how the parts of the glass and their organisation are responsible for it being fragile. One question is about what it is to be fragile for the glass and the other is about how it happens that the glass has the feature of fragility. Note that on both understandings, the constitutive questions are vastly different from causal questions. A causal question with regard to the fragility of the glas would ask "*how did the glass become fragile*?" or "*why did the glass break*?". Answers to this question would tell an aetiological story about how the glass broke or how it came to have the disposition to break. But the difference between the two non-causal but constitutive ways to take the question about the fragility of the glass translates to relevant differences between these two in terms of the explanantia that can be called upon given each understanding.

These approach-dependent differences in answering constitutive questions arise partly because identity conditions may not always be exhausted by statements about a system's parts and organisation. This may be true in cases in which the question itself does not concern a causal capacity. An example can be taken from Burge (2010, p. 379). In his view, one of the constitutive conditions of perception is that every perceptual state must have a veridicality condition. Having veridicality conditions, however, is not a matter of the parts and organisation of a system, but a condition that is an intellectual normative/epistemic property of perceptual states. Putting forward something's constitutive identity conditions is therefore not necessarily the same as providing a constitutive explanation in terms of a system's parts and organisational features, but it makes use of other kinds of explanation – here, normative/epistemic properties. Moreover if one accepts the possibility of multiple realisation of causal capacities, for example "being sighted", a statement about the parts and organisation of any specific types of a system's parts and their organisation instantiating this capacity may fall short of providing a list of conditions that would suit a constitutive answer in Burge's sense. In such cases, a statement on the parts or the organisation would fail to point out necessary conditions required for being sighted. A burgean answer presenting the identity conditions for the causal capacity would then have to provide identity conditions other than parts and organisations of systems, which – whatever they turned out to be – would have to be something beyond scope of the explanation used by constitutive explanations.

20 The most important points about constitutive explanations and the debates surrounding them have been usefully reviewed by Ylikoski (2013), on whose efforts I rely in the following.

Constitutive explanations are non-causal, and like most non-causal kinds of explanation, constitutive explanations have received relatively little philosophical attention compared to causal explanations. Relevant early work on constitutive explanation was done by Cummins (1975, 1983), complemented more recently by Craver (2007a). Constitutive explanations, as mentioned, aim to explain the causal capacities of a system (Cummins 1975, 1983; Harré and Madden 1975). These causal capacities are understood as the dispositions of a system to bring about a certain causally influential event or occurrence, given specific triggering or enabling conditions. In absence of these enabling conditions, causal capacities are powers existing unactualised in the system. These causal capacities can be explained by the parts of the system and their organisation, by virtue of which they are present in the system. Providing such an explanation, however, is not providing a causal explanation because it does not provide a causal story of *why* the system is doing what it is doing by spelling out the causal aetiology of its behaviour. Instead, the explanation accounts for *how* the system's components and their organisation instantiate the causal capacity to make this behaviour happen. As Cummins (2000, p. 122) lucidly sums it up: "The constitutive questions abstract away from the behavior and orchestrated activities of the parts, and ask how the system has a capacity for this kind of behavior."

This being the basic idea of a constitutive explanation, there are three further important peculiarities that I will also discuss in elaborating how diagnostic modelling is to be understood as constitutive. First, constitutive explanations are provided in what has been called a constitutive field; second, the scope of constitutive explanations usually entails single dispositional features; and third, constitutive factors used for an explanation may be used at various levels of description and grain. Let me start out with the first point, concerning the constitutive field. Now that we have an understanding of what constitutive explanations are, let me supplement this understanding with some ideas about *how* these explanations explain.

To understand how constitutive explanations explain, Ylikoski (2013) suggests combining two general approaches to explanation from philosophy of science: the *contrastive question approach* (e.g., Garfinkel, 1981; Hesslow, 1983; Lipton, 1991; Ylikoski, 2007; Craver, 2007b) and the *difference-maker* account (e.g. Mackie, 1974; Woodward, 2003; Waters, 2007; Strevens, 2008). The idea is that constitutive explanations have explanatory power because they treat their explanatory questions as *contrastive questions* that they answer by identifying the *difference-makers* responsible for the factual differences pointed out by the contrastive questions. What this means needs some explanation.

According to the contrastive question approach, the epistemic value of explanations is that they tell us why some fact rather than some other exclusive alternative fact (or group of facts) holds true. The exclusive alternative facts considered for this purpose are called the contrast class. When we ask an explanatory question, we do not always put forward a contrast class explicitly; rather, the contrast class is of-

ten implicit in the background. However, awareness of the assumed contrast class is central to clarity about what exactly it is that ought to be explained by an answer to an explanatory question. If one has such awareness, rather than only asking why X ϕ´s, the question is why X ϕ´s rather than X ψ´s, where ϕ and ψ are exclusive alternative facts about the subject X. The contrastive question approach assumes that all explanatory questions can be understood along these lines. Whether one believes this or not, let us for now follow Ylikoski in his claim that this understanding is at least useful for grasping constitutive explanations.

As an example, consider the explanatory question of why a certain animal species is found in the Atlantic Ocean. Although one might have an intuitive understanding what kind of explanation this question is asking for, it is actually ambiguous. It does not tell us what about the fact that this species is found in the Atlantic should be explained. It could be a) that it lives in this ocean rather than not living in it; b) that it lives in this ocean but not in *any* another ocean, or that it lives in this ocean but not in a specific other ocean; c) that it lives in this ocean but not on land as well; (and so on). Alternatively, the question could aim to address all these contrasts at the same time. Depending on which of these contrastive facts are considered to be part of the contrast class assumed for the question, the answer to the question will look very different. An explanation of why a species lives in in the Atlantic Ocean at all rather than not living there will obviously look different from explaining why it lives there but not also somewhere on dry land. Of course, in principle one can include every possible alternative fact in the contrast class. This would make matters incredibly complex, however, and turn a question that needs one answer into a question addressing multiple things that need to be answered separately at the same time. If a question is thought about clearly, it usually ends up having only one contrastive fact; indeed, a question may need dividing into a set of questions if it had more than one item in its contrast class before.[21]

Applied to constitutive explanation, the contrastive question framework considers the question asked to be why a system has a causal capacity opposed to alternative exclusive facts. Again, a precise explanatory question will have a specific difference that is intended to be explained, set by the chosen contrast class. If, for example,

21 As Ylikoski (2013) points out, we can also use the approach to reference classes to determine the difference between causal and constitutive questions: "Thus the contrastive thesis is not a claim about what people have in mind when they put forward an explanation-seeking question, but what they should have in mind (Ylikoski, 2007). Quite often the original scientific research question, when articulated in contrastive terms, turns out to be a whole set of related contrastive questions. This is a good thing: smaller questions are something we can actually hope to answer by means of scientific enquiry. And of course, nothing prevents one from asking both causal and constitutive questions – and questions driving scientific research are often such – but the contribution of the contrastive idea is to make it possible to analytically distinguish questions that require separate answers" (ibid., p. 123).

what should be explained constitutively is the fragility of a glass, this could be spelled out as the contrastive question for what makes a glass fragile, as opposed to the assumed contrast class. This contrast class may contain the property of robustness, so that a constitutive explanation has to answer which constitutive factors make the glass fragile rather than robust. Alternatively, the contrast class might contain the disposition to liquify under force. Whatever the contrast class looks like, it directs the explanatory effort to what should be constitutively explained about a present causal capacity.

The second component of how contrastive explanations explain is the difference-maker account. While the contrastive question account provides a better grasp on what the exact target of an explanation (its *explanandum*) is, the difference-maker account addresses what the explanatory components (the *explanans*) should be. The idea of this account is that a targeted fact or state of affairs is explained by pointing out what is responsible for this state or fact (as opposed to alternative states or facts) obtaining. The explanation thus identifies counterfactual dependencies, claiming that if the *explanans* had differed, the *explanandum* would have been different too. Combined with the contrastive question account, the task is then more precisely to identified counterfactual dependencies on responsible difference-makers for the targeted fact hold true, as opposed to the alternative facts contained in the contrast class. As Ylikoski (2013, p. 291) puts it:

> The idea of the explanans as the difference-maker is a natural partner of the idea of contrastive explanandum. Together they provide a powerful heuristic of scientific research. First, you create, find, or imagine the difference to be explained, and then you proceed to find the differences between the cases. Then you test whether these candidates can really make the difference, by testing whether they can bring about the difference to be explained.

In causal explanations, this would mean providing an understanding of the counterfactual dependency of an event Y on a previous event X, such that if X had not happened Y would not have occurred, given certain background conditions (the relevant constitutive field). Here X is the difference-maker causally explaining the presence of Y (Woodward 1984, 2003; Ylikoski and Kuorikoski, 2010).

The difference-maker approach also matches with constitutive explanation. Considered along these lines, providing the constitutive explanation means providing the conditions for the fact that a certain causal capacity rather than a chosen contrasting capacity or the absence of this capacity is realised in the system, by providing a statement about the parts and their organisation in a system (given certain constitutive field conditions) on whose presence the fact that the capacity in question is in fact present counterfactually depends. Constitutive explanations explain by providing such counterfactual conditions as difference-makers.

In the preceding paragraphs I have taken some time to introduce constitutive explanations in their structure and explanatory capacity. Now let me show how this fits together with modelling in general and diagnostic modelling in particular, in order to spell out the constitutive aspect of the constitutive-indicator strategy pursued by diagnostic modelling.

In general, for a model to qualify as a constitutive model, the model will be required to entertain the constitutive factors of a phenomenon in the model. More precisely, these constitutive factors, which make up the *explanans* of the constitutive explanation of a phenomenon, must be covered within the set cope of the model by being assigned to parts of the model structure, with the purpose of making the model a model of what would be targeted as an *explanandum* by the corresponding constitutive explanation. A model build based on a constitutive explanation of a phenomenon would therefore take the explanation as a background theory for setting up the model structure, deriving its plausibility as an adequate model of the phenomenon of interest from the plausibility of the background theory (the constitutive explanation) it is capitalising on. If the model were then used in a diagnostic manner – that is, to indicate the presence or absence of a constitutively modelled phenomenon in the targeted system as result of a model/world comparison – then the overall plausibility that this model could be used for such indicative purposes would rely on the quality of the constitutive explanation. If a constitutive explanation is assumed to capture the relevant constitutive factors of the disposition targeted by the model, it can be assumed that their presence indicates the model's target to be present, and thus that the model is a valid indicator. By using a constitutive model along these lines as an indicator, it also provides an explanation. More precisely, it provides a constitutive explanation in terms of difference-making. By pointing out the presence or absence of relevant constitutive factors that make the difference between the presence or absence of the disposition, the constitutive model tested in the model/world comparison provides a constitutive explanation by accepting or rejecting the compared model as adequate, and thus explains by saying that the factors that make the difference between absence and presence of the disposition apply or not.

Equipped with an overview of constitutive modelling and its indicative use in general, let us now turn to its use in the specific cases of diagnostic modelling. For this purpose, I will again speak of normative models as well as error models.

As established earlier (2.3), normative models give an idea of the diagnosed system in terms of its disposition to realise a certain causal capacity (output) given certain triggering conditions (input). The identification of a certain type of error in normative-model diagnostic reasoning takes place based on the type of abnormality that shows up in the comparison between the normative model and the actual behaviour of the system. According to the way it behaves differently from the normal input-output behaviour, the system can then be classified as presenting a certain

type of error. In the assumptions regarding the normative model, we therefore find a set of nested assumptions about what constitutes a normal system behaviour given an array of different inputs. In other words, the disposition to behave normally given certain triggering or input conditions is constituted by providing a certain output in response. Thus, the normative model is an amalgamation of a set of assumptions about what constitutes normal behaviour for the system in terms of inputs and outputs. The inapplicability of the model, given certain inputs, is then considered an indicator of the abnormal functioning of the system in one of the functions covered by it. The inapplicability thereby explains the assumption of the presence of the error constitutively by the absence of the constitutive factors that would render the system free form error.

In the case of error models, they too are structures geared towards matching up with constituents, but these models differ from normative models in two relevant regards. First, they are targeting constituents of the system's disposition that are considered to be responsible for an error occurring in the system, rather than those responsible for normal functioning. By identifying the presence of the constituents of a specific error, constitutive error models thereby explain the presence of the error by identifying the relevant difference-making factors responsible for it. And second, every error model is intended to target one specific instance of error to diagnose one specific error, and is thus informed by one constitutive explanation rather than being an amalgam of multiple explanations as in the case of a normative model. Their use as an indicator then works in differential diagnostics by testing their applicability in model-world compression with the erroneous system. If the presence of the specific set of constituents assumed in the error model can be found in the real world system this justifies the model's acceptance and thus justifies to apply the diagnostic label that the error the model is considered to be a model of. In the case of exclusion diagnostics, the inapplicability of any of these models – and therefore the inability to identify the relevant constituents for any condition that may be ascribed based on the application of error models – may be marked by a corresponding diagnostic labelling of the present error in the system that could not be modelled more specifically by an error model. Thus error models used in exclusion diagnostics explain by noting the absence of any set of relevant difference-makers that would indicate the presence of a certain error entity.[22]

22 In line with my remarks in footnote 18, where the discussed indication may rest on a causal or a correlational relationship, and noting how these options are mirrored in contemporary methodological debates in psychology, I want to point out something similar for constitutive indicators. While the idea of a constitutive indicator has to my knowledge not been discussed in detail in philosophy, it is present in the psychological research literature. There, the kind of indicator I am calling *constitutive* is called *composite*. Similarly to my discussion, Bollen and Bauldry (2011, p. 6) introduce them by saying that "Composite indicators are weighted elements that form a composite variable for which there is no disturbance term. That is,

Now that I have discussed the constitution aspect of modelling following the constitutive indicator strategy, I will now analyse what inferential patters are facilitated by this strategy. Doing so will be especially relevant to the last section of this chapter, since I will address the question of how qualitative diagnostic modelling justifies its outcomes by following the constitutive indicator strategy.

2.4.3 Inferential Patterns of Model-Based Diagnostics

In this section, I will discuss two inferential patterns occurring in the context of model-based constitutive indicator diagnostics. The first type of inference made in model-based diagnostics is *abduction* or inference to the best explanation (IBE) (2.4.3.1).[23] This type of inference occurs when the diagnostic modeller engages in basic normative-model diagnostics and in error-model diagnostics as forms of differential diagnostics. The second type of inference is *apophatic* inference (2.4.3.2). The name derives from ἀπόφημι (from *apophēmi*, "to deny") and is an adjective meaning "involving knowledge obtained by negation" (Harper, 2023). This type of inference, or rather (as I will discuss later) one of its specific instances, occurs if the normative model can be applied to the system and an error occurs, but as a result of the diagnostic evaluation none of the diagnostic error models for this error can be applied – in other words, in exclusion diagnostics.

---

the composite variable is an exact linear combination of the composite indicator variables. But beyond having no disturbance, the composite indicator coefficients are not structural or causal coefficients. Rather their coefficients are weights to apply to form the composite variable that is made up of them." Furthermore, they also bring up another point that was introduced in my discussion of constitutive explanations – namely, that constitutive explanations can address various aspects of the system that are relevant to realising a power of interest in the system on various level of description, a quality that plausibly carries over to constitutive models as well. As they put it, in discussion of social variables (e.g., character traits as variables): "Composite indicators do not necessarily have conceptual unity, but can be an arbitrary combination of variables" (ibid.).

23 Since my aim is to bring the theoretical considerations of this chapter to bear on psychiatric diagnostics, let me point out here that I am aware that the claim that IBE plays a role in medical diagnostics is not a new idea (e.g., Lipton, 1991; Console and Torasso, 1991; Gabbay and Woods, 2005; Aliseda and Leonides, 2013; Johnson, 2019; Stanley and Nyrup, 2020) and that it has also been raised with regard to psychiatry (e.g., Reznek, 1998; Vertue et al., 2008). What is unique about my account and what makes it preferable will be outlined in the final chapter.

Table 2: Types of inference (left) matched with inferential practices (right)

Type of inference	**Corresponding diagnostic practice**
Inferences to the best explanation (or abduction)	Normative-model diagnostics Error-model differential diagnostics
Apophatic inferences	Error-model exclusion diagnostics

2.4.3.1 Abduction or inference to the best explanation

Abduction as a reasoning pattern was introduced by Peirce (1878, 1903), in contrast to the two other widely discussed patterns of reasoning, induction and deduction. Deductions are *non-ampliative* (they do not add to what is already known) and *certain* (their conclusions must be true if their premises are true). Both inductions and abductions are *ampliative* and *uncertain*, which means that if true they extend our knowledge of the world, but that in contrast to deductions, their truth is not guaranteed by the truth of their premises. In inductions, properties or regularities are transferred from past events to the future, or from the observed to the unobserved. The difference between abduction and induction regards their target. While inductions aim to make inferences about future or unobserved events, abductions aim to infer something about the unobserved causes or explanatory reasons for an observed event (Aliseda, 2006). Since the distinctions amongst these three types of inferences were introduced, philosophers have recognised that abduction itself is not a single pattern of reasoning but consists of a collection of patterns of inference. One particularly valuable attempt to defend and systematise the various patterns of abduction or inference to the best explanation (IBE) is provided by Schurz (2008).

Following Schurz (2008), "the crucial function of a pattern of abduction or IBE consists in its function as a *search strategy* which leads us, for a given kind of *scenario*, in a reasonable time to a most promising explanatory conjecture test which is then subject to further test" (ibid., p. 205). This function can be fulfilled in different ways following different patterns of abductive reasoning. These different patters fall into two broad classes: *selective abductions* and *creative abductions*. In selective abduction, the task is to make a choice between competing alternatives that might explain the features the target phenomenon, while in creative abduction the task of the reasoner is to come up with a new explanation given the *explanandum* and potential constraints deriving from further circumstantial knowledge (ibid). Given my description of the inferential strategy of model-based diagnostics, I will concentrate on selective abductions as the one most plausibly at work in this form of diagnostic

practice. More specifically, I will concentrate on a class of abductions that Schurz calls *factual abductions*, and its subtype of *observable-fact abduction*.[24]

Factual abduction is the classic and most widely discussed form of abduction, introduced by the young Peirce (1878) himself (calling it *hypothesis*), before he generalised his understanding of abduction (Pierce, 1903) along the lines presented earlier. As Schurz (2008, pp. 207–208) puts it:

> In factual abductions, both the evidence to be explained and the abduced hypothesis are *singular facts*. Factual abductions are always *driven* by known implicational laws going from causes to effects, and the abduced hypotheses are found by backward reasoning, inverse to the direction of the lawlike implications. (…) It has the following structure (the double line === always indicates that the inference is uncertain and preliminary):
>
> (FA): *Known Law:* If Cx, then Ex
> *Known Evidence:* Ea has occurred
>
> ==========
>
> *Abduced Conjecture:* Ca could be the reason.

Observable-fact abduction is a sub-pattern of factual abduction. As Schurz argues, it occurs if there is a follow-up test procedure for the abduced conjecture such that "the follow-up test-procedure consists in the attempt to gain direct evidence for the abduced conjecture" (ibid., 207). Schurz offers the example of a murder investigation: "In the example of a murderer case, such direct evidence would be given, for example, by a confession of the putative murderer to have committed the crime" (ibid.).[25] Let us now see how the two inferential (sub)patterns of abduction, *factual abduction* and *observable-fact abduction*, apply to diagnostic modelling.

First, factual abduction applies to simple normative-model diagnostics. To recap, in this kind of diagnostic modelling, the modeller first recognises that the system produces an output that does not seem to be in line with its expected normal be-

24 Other forms of abduction irrelevant to my purposes but discussed elsewhere include law abduction, second- order existential abduction and its subtypes (micro-part abduction, analogical abduction, hypothetical cause abduction, speculative abduction), common-cause abduction and its subtypes (strict common-cause abduction, statistical factor analysis, abduction to reality), and theoretical-model abduction. If you are interested in these, I suggest you consult Schurz's excellent (2008) paper.

25 Philosophers such as Fumerton (1980, p. 592 f.), have claimed that abduction could be reduced to induction. While I am not able to discuss this claim here in detail, please see Schurz (2008, p. 207 f.) for a counterargument.

haviour. Then, to determine whether the system showing the *prima facie* erroneous output really is in error, the modeller has to ensure that their normative model really indicates a deviation of the system from normal behaviour in this situation. If this is the case, the system's behaviour can be classified as presenting an error. The inference taking place can then be mapped onto factual abduction. The modeller's background assumption is that a certain a kind of erroneous output of the system (Ex) usually occurs as a consequence of some (not further specified) alteration of the system (Cx), such that if there is a relevant sort of alteration (Cx) then the error occurs (Ex). That the error has occurred in the system (Ea) then justifies the inference that some error in the system is present (i.e., that Ca could be the reason).

The more specific subtype of IBE, *observable-fact abduction,* occurs in the case of error-model differential diagnostics. In this case, diagnostic inference is not based solely on a system's deviation from its behaviour as predicted by the normative model, suggesting some constitutive alterations in the system. In addition, and more specifically, error-model differential diagnostics takes place in the form of evaluation of the specific changes occurring in the system against specific error models. These models represent potential alterations of the system that may constitute the system's disposition to produce the error. Using these models can serve the diagnostic process in terms of differential diagnostics in two ways.

The first way for error models to serve differential diagnostics occurs if only one error model is known that should be applicable to the system if a certain error occurs. In this case, the error model may apply, and if so, the error model further supports the diagnosis provided based on the initial normative-model diagnosis, by showing that the specific setup of the system that is known to potentially bring about the error can indeed be found in the system. Alternatively the error model does not apply; therefore the diagnostic conclusion will be that the error initially identified with the aid of the normative model is present, but that it is an instance of the error not covered by the diagnostic understanding provided by the error model. This scenario turns the process into exclusion diagnostics, which will discussed in more detail below in connection with apophatic inferences.

The second way occurs if more than one potential error model exists that might match the system to explain the occurrence of the error beyond what could be said based on the normative model, and if indeed one of these models applies. In this case, the error found in the system can be identified as a specific instance of the initial error and can therefore be classified by a more specific diagnostic label. Again, it might turn out that no error model applies, which would, as before, lead to exclusion diagnostics, to be discussed in the context of apophatic inferences.

If we stick to the cases in which the modeller is successful in their attempt to apply an error model to the system, observable-fact abduction takes place. In this case, beyond the previously illustrated step of normative-model diagnostics and its factual abduction, an additional round of abduction following the same schema takes

place. This time, the modelling process does not take an erroneous output of the system as evidence; instead, it takes more specific features of the system as constitutive, whose application is supposed to indicate a more specific error than the rather abstract error attribution based on a normative model. This act of *observable-fact abduction*, looking for specific evidence to support a diagnostic claim, may thereby test the error model that would support the initial diagnosis of a normative-model based diagnostic conclusion. Alternatively, if the modeller's understanding of the system is more differentiated, such that multiple types of error might stand behind the error that would be recognised based solely on normative-model diagnostics, a differential diagnostic process would take place in which multiple error models would be applied to the system to make an observable fact abduction to the more specific error type they suggest. As mentioned earlier, if no model applies, the modeller may instead end up in an exclusion diagnostic process, involving apophatic inference, which I will discuss next.

2.4.3.2 Apophatic Inferences

Next in line are cases of what I will call *apophatic* inferences. Apophatic (from ἀπόφασις, to deny) inferences are not an inferential pattern in themselves, but they are instantiations of the commonly discussed inferential patterns (induction, deduction, and abduction) that draw *negative conclusions*. In philosophy, apophatic inferences have been discussed since Plato, and became especially prominent in Middle Platonism and Neoplatonism, via the still existing branch of theology called negative theology (Westerkamp, 2006).[26] For the analytic tradition, the idea of attaining knowledge by negative conclusions is also a familiar one, thanks to Popper's (1935) emphasis on falsification in the critical-rationalist approach. However, in recent years apophatic inferences have attracted attention mainly outside of the analytic tradition.[27]

Before I come to how apophatic inferences occur in model-based diagnostics, I shall begin by saying more about the nature of negative conclusions, their truth conditions, and their informational value. First of all, I will say something about how I will handle the most distinctive feature of apophatic inferences: negation. Negation in natural language and logic is a complex topic, a comprehensive treatment of which is beyond the scope of this chapter. For my purposes here, I will focus on

26 Please note that my use of the label *apophatic* does not suggest that there is a full match between the methodology of negative theology and the types of inference I describe here. I chose the label because I see a broad resemblance in the type of approach – namely, a pattern of inferences trying to arrive at an ultimate statement about a target by means of negative ascriptions.

27 Indeed it seems that the most recent debates in philosophical circles that have tried to actualise the idea of the *via negativa* have taken place among theologists and philosophers sympathetic to poststructuralist philosophy (e.g., Derrida, 1995; Ferretter, 2001; Rubenstein, 2003).

negations understood as indicative-mode declarations of negative predications, as originally discussed by Aristotle (De Interpretatione, 17a25). In other words, I will adopt the view that negations are statements consisting of a subject and a denied predicate applied to a subject that together form a proposition: C does not apply to a.[28] With this clarified, let me next discuss what kind of information we can gain from negative conclusions.

Prima facie, negative statements do not seem to correspond to specific facts that would serve as truthmakers of the negative statement in question. Rather, the informative value of such statements seems to lie within the information about the absence of the truthmakers of a state of affairs denied by the statement.[29] Therefore the informational content can be derived almost trivially from the negation itself: If the negative statement is true, it is not possible that any set of minimally sufficient facts from the set of all necessary and sufficient facts that would be truthmakers of the positive formulation of the negative statement hold true.

---

28 Why did I choose the Aristotelian model of negation considering negations to be part of propositions (*C* does not apply to *a*), rather than the model of negation from Fregean logic that considers negations to be denials of propositions ("it is not the case that *p(p=Ca)*")? The reason lies with the scope of the negation that makes each of these negations true. A negation expressed according to Fregean logic would be true under two circumstances: first if *C* does not apply to *a*, and second if there is no *a*. While the negation according to the Aristotelian logic claiming that *C* does not apply to *a* is true if there is an *a* and *C* does not apply to it, it is false if there is no subject on which the *C* could be predicated. Intuitively, these conditions make the Aristotelian model closer than the Fregean model to our natural language use of negation, and also closer to the use of negative statements in diagnostics. If I make a diagnostic statement that a certain error does not occur in a system, to claim that this statement is right since the system I am talking about does not exist seems strange. Rather than saying that this statement was right, it seems plausible to say that this statement is wrong or meaningless because the system I am talking about does not exist. Consequently, the Aristotelian model of negation seems more adequate to understanding negation in the context of system diagnostics. For a more in-depth discussion of Aristotle's understanding of negation and its defence against criticism from modern logicians, see Perälä (2020). For a comprehensive discussion of negation in natural language and logic in general, see, e.g., Horn and Wansing (2020).

29 In this point I basically side with Lewis (2001) approach to truthmakers in that I do not think there are specific facts that are truthmakers for everything that is (or can be) true. Rather, my view is that negations are true due to the fact that in current state of affairs, facts that would be truthmakers for the affirmative equivalent of the negation do not hold. This approach helps to avoid problems occurring if one begins to look for specific facts serving as truthmakers of negations, such as the so-called Paradox of Negation that concerns the questions "If a positive statement refers or corresponds to a positive fact, to what state of affairs does a negative statement refer or correspond?" and "What in fact is a negative fact?" (Horn et al., 2020).

Taking this for granted, it appears that information attained by a negative judgement as a conclusion of an inferential process is therefore relatively limited. This apparent poverty of negative statements was already pointed out by Plato in *The Sophist* when he stressed that it is in the nature of negative judgements to suffer from a lack of specificity, as all we learn from them is what is not the case, making them in general less informative than positive judgements.[30] This, however, is not strictly true. The informativeness of negative judgements depends partly on the context of their assertion, more precisely on the space of possibilities that forms the contrast class to the negative judgement. The relation is such that the smaller this contrast class is, the larger the informative value of a negative judgement becomes. Let us look at an example. If I make the negative judgement that my grandfather is not alive, this judgement has a high informational value given that the relevant contrast class of "being alive" contains only one alternative if we apply it to people who are already born, namely "being dead". The informational value is lower if, for example, I make the judgement that my father is not a bachelor, as the relevant contrast class to "being a bachelor" contains not one but several options. The man might be a fiancé, a spouse, or a divorcé. Even less informative would be the statement that something is not green, or, even worse, that something does not weigh 15 kg, since the intuitively chosen relevant contrast classes (i.e., all other colours or all other possible weights) form larger and larger contrast classes. From this it is clear that the scope of possible alternatives seems to determine the informative value of negative judgements. The claim that negative judgements are in *general* less informative than positive judgements has to be specified by saying that they are less informative as long as there is more than one alternative exclusive state of affairs, and they become the less informative the more such alternatives exist in the relevant contrast class.

Now that the informative value of apophatic inferences has been discussed, let me come to the relevant instantiation of apophatic inferences in model-based diagnostics. They occur as deductive inferences, instantiated as *modus tollens:*

(AI – D): *Known Law:* If Cx, then Ex
Known Evidence: Not Ea

Apophatic conclusion: C does not apply to a

Let us see how this applies to model-based diagnostics. Here such apotheic judgements occur if the behaviour of a system *prima facie* suggests a certain type of error

30 See Xenakis (1959) and Lee (1972) for a detailed treatment of Plato's thoughts on the informativeness of negative statements.

and a normative model is applied to the system – that is, normative-model diagnostics has taken place – but if, as a result of closer diagnostic evaluation based on more specific diagnostic error models, none of the tested models applied.

The attempt to apply these models to find the correct diagnosis is made with each of the models considered to suggest the presence of a certain kind of error (If Cx, then Ex / If Kx, then Lx/ …). However, if it turns out that none of the models applies (Not Ca/ Not Lx/ …), then none of the observed errors can be diagnosed (E, L, … does not apply to a). As a result, a finer-grained diagnostic judgement is not feasible. While the initial diagnostic evaluation *qua* abduction allows us to determine a type of error to be present, the second level of evaluation is based on more specific error models that add information about what potential instance of this error is not taking place in this system. The result is an instance of exclusion diagnostics.

So far, model-based diagnostics as introduced in the first half of this chapter has been elaborated regarding its inferential strategy and the inferential pattern at work in it. Presenting the inferential strategy has made clear the rationale for believing that this approach achieves its epistemic goals of correctly indicating and classifying errors in a system. Discussing the inferential patterns at work in model-based diagnostics has clarified what justificatory procedure is present in which aspect of model-based diagnostics. From this, I will now transition to the closely related question how we should think of the justificatory states of results obtained from a model-based diagnostic process.

2.4.4 Model-Based Diagnostics and Justification

To discuss the justification of conclusions in model-based diagnostics, I will distinguish between their internal epistemic justification (2.4.4.1) and external epistemic justification (2.4.4.2). Essentially, when I talk about the internal justification of diagnostic conclusion in model-based diagnostics, I mean the epistemic source of a justification a conclusion received within an assumed diagnostic system (a set of diagnostic models to diagnose a certain system). When I talk about external justification, I am referring to the source of justification that is outside the system insofar as it provides reason to trust the framework of a diagnostic system used for diagnostic modelling in the first place. To put it briefly: internal justification is concerned with the source of epistemic justification *within* the diagnostic procedure, while external justification deals with the diagnostic justification *of* the diagnostic procedure. Let me expand a little more on both types of justification to ensure that the difference is clear.

In internal justification, the justification enjoyed by the diagnostic conclusion within an adopted diagnostic system of model-based diagnostics is achieved by virtue of meeting the internal standard assumed by model-based diagnostics as a strategy to arrive at its diagnostic conclusions. By *internal standards* I mean the

epistemic norms of diagnostic procedure that a given diagnostic system needs to follow in order to arrive at a diagnostic conclusion considered adequate within this framework. Looking at internal justification allows us to identify, for example, how conclusions within an established diagnostic system come to be deemed justified. As I will argue, the epistemic core value relevant to internal justification in diagnostic modelling is reliability.

External justification, on the other, is the justification a diagnostic conclusion enjoys by virtue of being a product of an epistemic procedure meeting the "epistemic gold standard" (Schurz, 2011) – namely, being truth-conducive.[31] Whether the epistemic procedure of diagnostic modelling meets this gold standard will not depend on the plausible internal framework used to justify its conclusions, but will rather be based on how good our reasons are for claiming that the procedure that is producing results is indeed producing correct results. In other words, the question is whether a diagnostic modelling process is following the general approach and employing a certain set of diagnostic models to diagnose a system in a way that is reliable, and of which we also have reason to believe that its outcomes track actual instances of errors in the system. What is at stake here is the validity of a given modelling procedure. As I will argue, this validity depends on the quality of constitutive explanations that are used to infer the absence or presence of certain error conditions.

Discussion of the justification of conclusions in model-based diagnostics is crucial to for allowing us to address the Methodological Question. It is crucial because in a methodology we want to understand how a method justifies. Discussing internal and external justification separately for this purpose is important to ensure that the considered method follows an internally rational route to come to conclusions that we can make comprehensive in a theory of this method (internal justification). Beyond being internally comprehensive, it is also important that we have reason to believe that a method performs well in its application to the real world and that we should trust its results, or at least that we know to what extent we can trust its results (external justification). I will begin by addressing internal justification.

2.4.4.1 Internal Justification

To address the internal justification of diagnostic conclusions, let me quickly review some aspects of the model-based approach. Diagnostic modelling follows the constitutive indicator strategy. In brief, this means that diagnostic conclusions in a given diagnostic system are drawn by testing the (in)applicability of normative and error models. The results of these comparisons are then used in different ways (normative-model diagnostics, error-model diagnostics) to indicate the absence or presence of errors. Since the occurrence of a (mis)match of a model used in the diagnostic

31 For a brief discussion of this standard view, see footnote 17, chapter 2.

process is taken to be an indicator, and since this indication is based on reliable correlation, our trust in the results rests on their justification *qua* the epistemic value of *reliability*. To bring an example to mind that highlights the centrality of reliability in the context of indication, think again of the example of the doorbell and the barking dog discussed in my earlier analysis of indication. What makes the barking dog a good indicator that the doorbell rang is that the dog almost always barks when the doorbell rings, and rarely barks on any other occasion. The barking is a good indicator because of its reliability. The same is true for diagnostic models: they are thought to be good indicators because of the reliable correlation of their (mis)match with the targeted system in case of the presence of the error they are intended to indicate. While we can thus say that reliability is crucial to the internal justification of diagnostic conclusions, one may expect there to be more to say about this. More specifically, one may hold the *prima facie* plausible intuition that the strength of internal justification for diagnostic conclusion in model-based diagnostics may depend on the type of inferential procedure used to produce it. Let me elaborate why one may think so.

One may think that although all inferential patterns used in model-based diagnostics rests on the justification by reliability, some of these patterns may provide better justification to conclusions than others. Should we not expect, for example, that error-model diagnostics would be better justified than normative-model diagnostics, given that, as we discussed earlier, error models assume a far more detailed understanding of specific errors that must be found present in diagnostic systems to allow for diagnostic conclusions, compared to the rather abstract assumptions of the normative model? This rhetorical question may sound *prima facie* plausible. One may reason along the following lines: the more details in a model that need to be assessed, the harder it is for the model to be fulfilled by a targeted system, so that conclusions that require a specific outcome in the assessment of a diagnostic model that is more detailed are harder to come by. If they are harder to come by, meeting these more demanding conditions should be assumed to provide better justification. However, on closer investigation this reasoning is wrong. What such reasoning actually tracks is not the internal justification of conclusions but their *informational value*, which as I will argue is not a source of intrinsic justification, since diagnostic modelling rests explicitly on indication – that is, on reliability. But before I argue along these lines, let us make clearer what I mean by informational value.

By the informational value of diagnostic conclusions, I mean the number of insights we have into a system based on a diagnosis given to it. Hence the informational value of a diagnosis equals the number of constitutive factors assumed in a diagnostic model that need to be matched with the modelled system to support a diagnostic conclusion about it. Let me give an example. Consider a certain portion of a normative model that assumes a normally functioning system to operate, so that it provides a certain output given a certain input, considers an error to be present

based solely on inputs and outputs to the system. Hence the informational value of a diagnosis based on a normative model will have the informational value of precisely this aspect of the system's behaviour: the erroneous input/output. Now compare this to a diagnosis based on an error model. An error model in the context of a differential diagnostic process consists of several propositions regarding constitutive facts required to be true about the system (beyond providing a certain output given some input), that must be present in the system for the model's successful application and therefore the justification of a diagnosis based on the model. The enabled error diagnosis in this case will be more informative than a diagnosis based on a normative model, since the error model goes beyond the normative model and provides further details to assess when making a diagnostic ascription to the system. So much for informational value and why it is higher in some approaches within model-based diagnostics than in others. Now we can return to the question of whether informational value provides internal justification, and thus whether some diagnostic conclusions have better intrinsic justification than others – in our example. Let me first explain why I believe informational value does not contribute to intrinsic justification.

The informational value of a diagnosis differs based on the modelling procedure enabling it, but the differences in informational value do not translate into differences of intrinsic justification. The intrinsic justification for thinking that a given diagnosis indicates a specific error rests on the assumption that the model used for the diagnosis allows us to reliably predict the presence of the error, hence the vehicle of intrinsic justification is indication, which is constituted by a reliability relationship. This reliability, however, does not depend on the informational value of a diagnosis, hence it is justified by the reliability and not the informativeness of the diagnosis. To illustrate this, let me return one last time to the example of the barking dog and the doorbell that I used earlier when explaining indication.

This time, imagine that we have two scenarios. In the first scenario, the dog barks whenever the doorbell is rung and not when it is not rung, but it makes no difference how often or how fast the bell is rung. In the second scenario, the dog barks only if the doorbell is rung twice and not when the doorbell is not rung or is rung more or fewer times. In both cases, the dog's barking reliably correlates with a state of affairs; therefore in both cases it indicates this state of affairs. However, the two states of affairs correlating with the barking of the dog differ specificity. In the first instance, the dog barks whenever the bell rings, and in the second, there is a specific pattern of ringing whenever the dog barks. It appears that if in both scenarios we cannot hear the ringing of the bell but only the barking, and if we are familiar with the barking behaviour of our dog, we would have more detailed knowledge about the obtaining state of affairs in the second scenario than in the first. In the second scenario, the barking of the dog indicates not only that it has rung, but more specifically that it has rung exactly twice, while the barking in the first scenario may indicate any num-

ber of rings. The barking of the dog in the second scenario therefore seems to be an indicator with more informational value. We know the bell rang, and we know it did so in a specific way: exactly twice. In the first scenario, we only know that the bell rung. However, assuming that the dog's barking is indeed reliably correlated with the relevant doorbell-ringing scenarios, it appears that our reasons to believe what both barks indicate are equally well justified in both scenarios, since both instances of barking indicate what they indicate with the same reliability. Because reliability understood as correlation is the determining factor of indication, this is all that counts in this context, no matter how informative the state of affairs may be that is indicated.

If we bring this back to modelling, we may think of modelling in analogy to these scenarios. The first scenario, in which the dog's barking provides only the information that the bell has rung, may be compared with a diagnosis based on the application of a normative model, providing a rather thin understanding of the presence of an error, only in terms of input and output patterns. The second barking scenario may be thought of in terms of error models used in differential diagnostics, since it indicates not only that the bell rang, but moreover that it was rung twice. By analogy, the use of an error model to provide a diagnosis goes beyond the assessment of abnormal input and output patterns and takes into account more specific aspects of the erroneous occurrence. Just as the reliable correlational relationship between barking and the state of affairs it indicates is what allows the instances of barking to justify the belief in the state of affairs indicated by them, it is the reliable relationship between the applicability of a model and the state of affairs (the error) it indicates that provides the resulting diagnosis with intrinsic justification. Just as the barking of the one dog is not better justified than the barking of the other because of informativeness but solely because of reliability, likewise a diagnosis based on one approach to diagnostic modelling is not better justified than one provided by another because of the informativeness about a state of affairs in the model based on the respective dogs barking. Given the nature of the indication relationship consisting in reliable correlation and given the fact that, as laid out in earlier sections, diagnostic modelling is supposed to use models as indicators for the presence of errors, the intrinsic justification of diagnostic conclusions is based on the vehicle of indication and thus on reliability as the central epistemic value.

While reliability is crucial for internal justification of diagnostic conclusions, it is not the whole story regarding justification. For internal justification, as presented here, to bear any general epistemic weight, we have to be convinced that a diagnostic system in model-based diagnostics that is able to justify things internally is also justified on a more fundamental level. We must be convinced that it not only provides us with an epistemically plausible way to think about conclusions as being justified within the system, but that the framework itself based on which these inferences are made is valid. The diagnostic system requires external justification. We may well

have a diagnostic system being used in model-based diagnostics that provides us with reliable (i.e., repeatable/stable) results, but we also need a reason to be sure that these results indeed track the presence of actual errors, a reason to believe that the outcomes are valid – a reason to believe that they really identify the presence of specific errors.[32] Showing how a diagnostic system (a set of diagnostic models used to address a certain system) in model-based diagnostics gains external justification is my next step.

2.4.4.2 External Justification

As discussed in detail in section 2.3, the basic approach of diagnostic modelling is to apply diagnostic models (normative models or error models) to a diagnosed system and use the result of the comparison procedure. The results are the identifiers of matches and mismatches between these models and the real-world system, and they are used as indicators for the presence of suspected errors. Therefore, as discussed in the previous subsection, the internal justification of diagnostic conclusions in diagnostic modelling rests on the assumption of the reliability with which the use of these models allows us to indicate the presence of the targeted errors. However, as I mentioned repeatedly, to assess whether the modelling used indeed reliably indicates a targeted phenomenon (i.e., a specific error), we need an additional source of justification. We need some external justification that provides us with reason to believe in the diagnostic results by ensuring the validity of the models that are used for purpose of indication. This would mean, for example, justifying that an error model indeed contains relevant constitutive factors of an error. Only then does the model seem a legitimate basis for an inference to best explanation regarding the presence of this error in a differential diagnostics procedure, hence making it permissible to use its applicability as an indicator for the presence of this error. If the need for this additional dimension of justification is acknowledged, the questions becomes: how do we gain this external justification for the validity of diagnostic models so that they can be assumed to be valid tools for use in the inferential machinery of diagnostic modelling producing externally justified diagnostic conclusions?

I argue that the external justification of diagnostic conclusion *qua* use of valid models in the diagnostic process depends on the justificatory strength of the background theories used to set up diagnostic models. To show why, we must compare the standard approach of modelling as presented in section 2.1 and the more specific use of modelling for diagnostic purposes. As discussed in section 2.1, in the attempt to develop a model of a system that allows for matching and simulation of

32 The relationship between reliability and validity that I presuppose here is the commonsensical understanding of the relationship between reliability and validity in measurement. Reliability of a measurement depends on its validity, whereas validity does not depend on reliability, and a valid measure is generally reliable (Bajpai and Bajpai, 2014).

certain features of a system, a modeller can use a range of sources to inspire their model structure. Modellers may be inspired by their intuition, draw on other models whose structure they deem promising to target the intended system's features, or capitalise on pre-existing theories of the phenomenon being modelled. A model structure derived from these sources will then usually be tested against the targeted real-world system in model/world comparison to see whether the model matches the real-world system well enough in terms of representational and dynamic fidelity. If not, the model will be revised until it does. Having a model successfully go through this process seem to justify the assumption that the model is accurate enough for the purposes of the modeller to be accepted as a model of the targeted system. If we look at the process of diagnostic modelling, however, there seem to be some differences.

Diagnostic modelling differs in some regards from the standard procedure for modelling a system that I have just sketched out. It does by virtue of its epistemic purpose. While modelling as just described aims to provide a good representation of the targeted system, diagnostic modelling attempts to make a judgement about the targeted system. The former kind of modelling takes the real-world system as a benchmark for the model and thus derives its legitimacy as a model by displaying model/world comparison. Diagnostic modelling, by contrast, takes its models as benchmarks to test the real-world system regarding features that suggest a diagnostic conclusion. If the diagnostic modelling process itself cannot equip the models used within it with plausibility, but requires their legitimacy before they are applied, then there are arguably ways to ground trust in these models.

The ways for diagnostic models to claim validity comes down to two options. The first option is that in the step of model construal, the diagnostic model is set up based on a background theory that provides a constitutive explanation either of what the normative behaviour of the system should look like (normative model) or of the constitutive factors of specific instances of error in a system (error model). The justification of the assumption that these models indeed capture the relevant constitutive factors of the system if a diagnosed condition is present then itself depends on the quality of the theories from which these models are derived. However, the question of when the acceptance of a theory is justified is in itself a highly controversial question that I will not be able to explore; what I can do here is merely to clarify that this is where the burden of justification shifts to. The second option is that the diagnostic models used are themselves results of earlier modelling attempts that were not diagnostic modelling, but rather system modelling, either focusing on normal system behaviour and its constituents in order to develop models for it or else aiming to provide such models for errors or abnormally behaving systems. Once these models have been developed, they can be reused by diagnostic modellers for their own purposes. As we will see in the next chapter, in psychiatry most diagnostic models rest either on our folk-psychological theories of human psychology and be-

haviour or on psychopathological research. For now, however, let us consider what is implied by these two ways for diagnostic model to gain plausibility.

If diagnostic models are derived either from theories providing the relevant constitutive explanations or from pre-established constitutive models, diagnostic conclusions enabled by the use of these diagnostic models can enjoy external justification. Diagnostic models that are used in model-based diagnostic procedures can enjoy support that grants them plausibility, so that their application to identify erroneous conditions seem to be justifiable. I say "seems", because these models will of course be only as plausible as the theory they are derived from or the quality of the modelling process that produced them. However, if these sources suffice, which is an empirical matter that would have to be evaluated for any given diagnostic system, it seems that diagnostic conclusions arrived at by the diagnostic modelling enjoy external justification. They enjoy external justification partly because the conclusions drawn within the system are a *supposedly* reliable way to become aware of errors in the system. But indeed, we have reason to believe that these inferences are also reliable in that they are based on adequate diagnostic models that are able to identify sufficiently relevant constituents of a targeted error, which in the presence of these constituents *qua* inferences to the best explanation allow us to take the applicability of these models as an indicator of the presence of the relevant error. External justification – that is, the validity of the models to be used as reliable indicators of error – would thus rest on the quality of the background theory and modelling approaches used to come up with the diagnostic models in the first place. Thus, the principal source of external justification for diagnostic conclusions is now laid out.[33]

33 This idea that the validity of diagnostic models depends on their capacity to pick out the constitutively relevant aspects of systems, enabling them to actually pin an error label to an underlying constitutively responsible makeup in the diagnosed system, is related to the understanding of validity in of test instruments. Test instruments are usually judged to be valid when they actually measure the construct at stake – a use of the notion going back at least to Kelley (1927). When is this "actually measuring" requirement satisfied? There are causal as well as correlational proposals. Borsboom and other psychometricians (Borsboom, Mellenbergh, and Van Heerden, 2004; Borsboom, 2005) proposed that a measuring procedure "is valid for measuring an attribute if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure" (Boorsboom, Mellenbergh, and Van Heerden, p. 1061). Many philosophers (e.g., Angner, 2011; Cartwright and Bradburn, 2001; Alexandrova, 2017; Michel, 2019), however, substitute b) for a mere correlational criterion. My understanding of validity would fit well with a correlational proposal. Note, however, that I do not mean to claim that model-based diagnostics is a measurement process; to evaluate whether this is true or not would require work that is beyond the scope of my project. All I was interested in here is giving a better grasp on the idea of validity.

2.5 Conclusion

In this chapter I have presented my account of models and modelling, which in its application to psychiatric diagnostics will provide my answer to the Methodological Question: What is the method of psychiatric diagnostics? Namely, that psychiatric diagnostic reasoning is qualitative, constitutive diagnostic modelling. The content of this chapter enables us to understand what qualitative, constitutive diagnostic modelling is and will provide the basis on which to formulate what needs to be shown about diagnostic reasoning to make plausible that it embodies this kind of modelling. Doing so will be the task of the next chapter. To end this chapter, let us briefly review what has been done.

In this chapter, I first introduced a general understanding of modelling as a process, understood as what is called the indirect strategy of representation. Next, I presented some specifications of modelling: one specification regarding a potential format of modelling, qualitative modelling, and one specification regarding a goal-driven epistemic approach one may take when using modelling – namely, diagnostic modelling. Thinking in terms of the Methodological Question, these parts of the chapter provided the description of the method I claim to be used in psychiatric diagnostics. Next, I discussed the inferential strategy pursued by diagnostic modelling, the constitutive indicator strategy, followed by an exploration of the inferential patterns that underlie the inferences generated via this route: inferences to the best explanation and apophatic inferences. Finally, I discussed the justification of conclusions drawn in model-based diagnostics. Again, thinking of the Methodological Question, this second part of the chapter provided material with which to address its remaining aspects, beyond the task of *describing* the method in place in diagnostic reasoning. The second half of this chapter provided the rationale behind the procedures of model-based diagnostics as a method as well as an understanding of how its conclusions are supposed to be deemed justified.

With all three aspects of an answer to the Methodological Question (description, rationale, and justification) in hand at the end of this chapter, the remaining task is to show that the method of diagnostic reasoning is indeed at work in diagnostic psychiatric reasoning and thus that the methodology presented here applies to it. This brings us to the next chapter, in which this task will be completed.