Chapter 3
Ockham's Proportionality: A Model Selection Criterion for Levels of Explanation

Jun Otsuka

Abstract Philosophers have long argued that a good explanation must describe its explanans at an appropriate level. This is particularly the case in social sciences and risk analyses, where phenomena of interest are often determined by both macro and micro factors. In the context of the interventionist account of causal explanation, Woodward (Philosophy of Science 81:691–713, [18]) has recently proposed that a cause must be proportional in the sense that it contains just enough information about its effect. The precise formulation of proportionality and its justification, however, have been under debate. This article proposes an interpretation of proportionality based on Akaike Information Criterion, a statistical technique for model selection. In a nutshell, disproportional cause variables with too much detail often call for extra parameters, which increases a model's complexity and impairs its predictive performance. By focusing on a model's predictive ability and its relationship to evidence, this chapter highlights the importance of a pragmatic or what Woodward calls a "functional" factor in the reductionism debate.

Keywords Ockham's razor · Reduction · Individualism vs holism · Model selection · Akaike information criterion

3.1 Introduction

Philosophers have long argued that a good explanation must not only identify the right explanans but also describe it at an appropriate *level* [2, 11, 20]. This is particularly the case in social sciences and risk analyses, where phenomena of interest are often determined by both macro and micro factors. The cause of poverty, for example, may be attributed on the hand to macrosociological factors such as recession, taxation system, the extent of the social safety net, etc., and on the other hand to individual characters such as job skills, education, or health status. The ubiquity of competing sociological theories of different granularity has raised a long-standing

J. Otsuka (✉)
Kyoto University, Yoshida-Hommachi, Sakyo-ku, Kyoto, Japan
e-mail: junotk@gmail.com

T. Matsuda et al. (eds.), *Risks and Regulation of New Technologies*,
Kobe University Monograph Series in Social Science Research,
https://doi.org/10.1007/978-981-15-8689-7_3

debate between individualists who try to understand any social phenomena in terms of behavior, properties, or interactions of individual actors, and holists who confer genuine explanatory roles on social structures or organizations [7, 9, 21]. The key question here is whether macro variables have any causal or explanatory power irreducible to properties of its parts, despite the fact that the former supervene on and thus are completely determined from the latter.

In the context of the interventionist account of causal explanation, Woodward [17] has recently proposed that a cause must be *proportional*, meaning that it must contain just enough information about its effect. This invites two questions: how to assess or measure proportionality, and why is proportionality a good thing? This article proposes an interpretation of proportionality based on Akaike Information Criterion (AIC; [1]). Akaike's theory tells that, other things being equal, predictions of parsimonious models tend to be more accurate than those of complex models [3]. Applying this idea, I will argue that disproportional cause variables with too much detail often call for extra parameters, which increase a model's complexity and impair its predictive performance. The proportionality criterion in this understanding is thus a variant of Ockham's razor applied to the context of causal explanations [14, 15].

The chapter unfolds as follows. I begin in Sect. 3.2 with a brief description of Woodward's notion of proportionality, followed by an examination of criticisms and interpretations of the concept offered by subsequent philosophical works (e.g., [4, 10]). Section 3.3 introduces my account of proportionality based on Akaike's theory. After its formulation, the idea will be illustrated with a simple simulation to compare the predictive accuracy of two—proportional and disproportional—models. The new approach for selecting a level of explanation has implications for reductionism, which are discussed in Sects. 3.4 and 3.5. The AIC-based proportionality clarifies conditions under which multiple realizability does not bar reductive explanations: in short, successful reduction occurs when a lower-level theory integrates micro-level properties into a simple model. The approach also highlights pragmatic factors in the reductionism debate, most notably our ability to collect data, as a key to deriving the positive value of higher-level explanations. I will argue this pragmatic nature makes my account of proportionality more in line with Woodward's [18] *functional approach* to explanations.

3.2 Kinds of Proportionality

Proportionality is the requirement that a description of a cause must "fit with" or "proportional" to that of an effect in the sense that it does not contain irrelevant detail. Consider Yablo's [20] example of a pigeon trained to peck at red targets to the exclusion of other colors. Now, suppose the red target the pigeon pecked on an occasion had a particular shade of scarlet. We then seem to have two ways of describing the situation:

1. The presentation of a red target caused the pigeon to peck.

2. The presentation of a scarlet target caused the pigeon to peck.

Provided they are both true, (1) strikes us to be a better explanation than (2) because by assumption what makes a difference in the pigeon's behavior is redness rather than scarletness. Proportionality captures this intuition. According to Woodward's definition, a cause is proportional to its effect iff (a) it explicitly or implicitly conveys accurate information about the conditions under which alternative states of the effect will be realized and (b) it conveys only such information—that is, the cause is not characterized in such a way that alternative states of it fail to be associated with changes in the effect ([17], p. 298).

In Yablo's example, describing the causative target as scarlet rather than red violates the second condition (b), because other non-scarlet reds, say dark red or rose, would still trigger the same pecking behavior, and thus these "alternative states fail to be associated with changes in the effect." Proportionality is devised to rule out such redundant information that plays no explanatory role.

Woodward's proposal has come under close scrutiny in recent philosophical discussions. Franklin-Hall [4] interprets proportionality as a requirement that the functional relationship between a cause and an effect be *bijective*—the first part (a) of the definition requiring each cause to be mapped to a specific effect (one cause, one effect), while the second part (b) forbidding distinct causes to be mapped to the same effect (one effect, one cause). Understood in this way, however, Franklin-Hall contends that proportionality fails to reject an intuitively too fine-grained explanation. She notes that the above descriptions (1) and (2) of Yablo's thought experiment are incomplete, because they do not specify the contrast class, i.e., what values the cause variable could take other than red (or scarlet). Franklin-Hall fills in that missing information and comes up with the following contrast class:

1* The presentation of a red target (other value: presentation of a non-red target) caused the pigeon to peck (other value: not peck).
2* The presentation of a scarlet target (other value: presentation of a cyan target) caused the pigeon to peck (other value: not peck).
([4], p. 564, with the order reversed).

Intuitively (1*) is the better explanation for the same reason we favored (1) above, but Franklin-Hall argues that proportionality fails to support this intuition because the causal relationships in (1*) and (2*) are both bijective: in (1*) we have {red → peck, nonred → notpeck}, while in (2*) {scarlet → peck, cyan → notpeck}.

This criticism, however, is an artifact of restricting the domain of the mapping relation to an arbitrary subset of all the target chips, which presumably include those that are neither scarlet nor cyan. What if samples contain, say, cobalt or navy targets? (2*) says nothing about their consequences and thus is at best an incomplete description of the causal relationship.[1] This could be patched by adding a third

[1] On the other hand, if indeed all targets are either scarlet or cyan, there is no difference in granularity between the two descriptions and choosing between them is simply a matter of taste. Note the problem here (when there are more than scarlet or cyan targets) is that Franklin-Hall's "variable" having only scarlet and cyan as values fails to satisfy a formal requirement of a random variable,

catch-all value such as "presentation of a target neither scarlet nor cyan," but then the domain has three causal values and the relationship is no longer proportional in the bijective sense.

Another—more sympathetic—interpretation comes from the group of Paul Griffiths and his collaborators, who use information theory to refine the concept of information in Woodward's definition of proportionality [5, 10]. Recall proportionality requires a cause X to convey enough information about the effect Y but no further. Griffiths et al. identify the amount of information that X carries about Y with their *mutual information* $I(X; Y)$ which represents the extent to which knowing a state of X reduces the uncertainty of Y. In contrast, the excess of information in cause X can be measured by its *entropy* $H(X)$ which represents the uncertainty about X's state. These two measures set conflicting objectives because fine-graining a variable increases both its mutual information (with any other variables, including its effect) and entropy. Proportionality can be defined as an optimal balance between these two desiderata:

Prop_{INF}: a cause X of an effect Y must (a′) maximize the mutual information $I(X; Y)$ while (b′) minimizing its entropy $H(X)$ [10].

As can be shown easily, this is equivalent to choosing the coarsest cause variable that maximizes the mutual information.

One strength of the information-theoretic interpretation is that it can handle continuous or stochastic variables. Suppose, as is very likely, that the pigeon in the above hypothetical experiment responds to stimuli only stochastically. Such a stochastic causal relationship cannot be expressed by a simple bijective function, but Prop_{INF} is applicable as long as we have the joint probability distribution over the cause and effect variables. In effect, relationships do not even have to be causal—one can well calculate Prop_{INF} for a correlational relationship with no direct causal link, although the focus of [10] is on causation. This holds true of any other proposal of proportionality, including Woodward's, Franklin-Hall's, and mine, and for this reason what follows treats proportionality as a criterion for the general problem of variable selection, not just for causes or effects.

Although theoretically attractive and versatile as seen above, the information-theoretic criterion is difficult to apply in actual problems because the knowledge of the joint probability distribution it requires is hard to come by. A pigeon's pecking probability, for example, is not something that is given a priori, but must be estimated from data (that is the reason we do experiments). Mutual information and entropy can also be calculated from data, but the problem is that the sample mutual information tends to *overfit* data. The assumption of Prop_{INF} is that mutual information hits a "plateau" as the causal variable gets fine-grained—the proportional variable is the coarsest among those at the plateau. But as we will see later with a simulation study, sample mutual information tends to increase almost indefinitely in proportion to the granularity of the used variable. This suggests Prop_{INF} is likely to fail to screen out too fine-grained descriptions in actual cases.

defined as a function on the sample space. Hence her later consideration on "exhaustivity" which amounts to adding *other* cause variables does not affect the argument here.

What motivates Woodward's account of proportionality (along with his other criteria, such as specificity) is what he calls the functional approach to causation, which evaluates causal claims in terms of their usefulness or functionality in achieving our epistemic goals and purposes [18]. The project in this line involves "*normative* assessment (and not just description) of various patterns of causal reasoning, of the usefulness of different causal concepts, and of procedures for relating causal claims to evidence" (p. 694, italics in original). The philosophical analysis of proportionality, then, must identify the specific epistemic goal it is supposed to serve and clarify its connection to evidence. That is, why, how, and when is proportionality a good thing? My proposal is that a proportional cause variable is expected to give more *accurate predictions* than non-proportional ones, in the case predictions are based on finite data. Proportionality, therefore, is not an a priori goal but rather a means to achive predictive accuracy, and the decision as to whether a given description is proportional or not depends not only on the nature of the causal relationship but also on the amount of data we have to estimate the relationship. The next section substantiates this idea based on Akaike's theory of model selection.

3.3 Model Selection Approach

The previous discussions on proportionality have asked what the appropriate level of description of a causal relationship is, assuming the relationship itself is already known. This assumption, however, is unrealistic because in most empirical research scientists have to begin by hypothesizing the relationship between a putative cause and effect. The hypothesized relationship is called a *model* and is represented by a function that calculates the probability of an effect given the input of a cause. In our pigeon example, a model assigns the probability of pecking to each target presented. There are various ways to model the same phenomenon. To illustrate this imagine two experimenters, Simplicio and Complicatio, come up with different models about the pigeon's behavior. Simplicio thinks the only thing that makes a difference in the pigeon's behavior is whether the target is RED or BLUE. He thus builds a model with two parameters which specify the probability of pecking targets of each color, $P(\text{peck}|\text{RED})$ and $P(\text{peck}|\text{BLUE})$. Complicatio thinks that's not enough. His hypothesis is that pigeons have better vision than human and can distinguish subtle nuances in color. Accordingly, the target chips that look red for us must be further classified into DARK RED, SCARLET, and ROSE, whereas the blue targets into CYAN, COBALT, and NAVY. Complicatio's model thus has six parameters, one for each conditional probability given a specific shade. This model is clearly more fine-grained than Simplicio's, and this difference in granularity is reflected in the number of parameters of the respective models.

To decide which model is better, they jointly run an experiment and fit their models to the obtained data. How well a model fits to data can be evaluated by looking at its *likelihood*, which is the probability of data given a model $P(\text{data}|M)$. High likelihood means that the observed data are well predicted by a model, which certainly seems a good sign. Since a model's likelihood depends on its parameters, one can choose the best set of parameters that maximizes a model's likelihood, or log-likelihood, which comes to be the same thing (taking the logarithm is just to make the calculation easier). Such parameters are called *maximum likelihood estimators*. In our case, they are actual frequencies of pecking—hence if pigeons have pecked 4 out of 10 total RED target presentations, the maximum likelihood estimator of $P(\text{peck}|\text{RED})$ is simply 0.4.

Suppose Simplicio and Complicatio have done their math and obtained the maximum likelihood of their model. Which model fits the data better? Without exception, the winner is Complicatio. In nested models like those we have here, the likelihood can only increase but never decrease as a model's parameters increase, because a model with more parameters is more flexible to "fit" the data, and this is so even if it contains seemingly redundant or unnecessary parameters. To borrow Hitchcock and Sober [6] expression, likelihood measures how well a model *accommodates* data, i.e., the facts that have already happened. In our case, Simplicio's model is a special case of Complicatio's with $P(\text{peck}|\text{DARKRED}) = P(\text{peck}|\text{SCARLET}) = P(\text{peck}|\text{ROSE})$ and $P(\text{peck}|\text{CYAN}) = P(\text{peck}|\text{COBALT}) = P(\text{peck}|\text{NAVY})$. This means any data that can be "accommodated" by Simplicio's model can be equally well handled by Complicatio's. This is a general phenomenon: any reductive model has a higher likelihood than its less-specific counterparts, and thus better accommodates data. This fact underlies the reductionist intuition that a lower-level description allows for a finer representation of the reality, and thus is epistemologically superior.

Accommodating the past, however, is not always our epistemic goal, nor is it even an important one. Hitchcock and Sober [6] rather emphasize *prediction* as a major goal of building scientific models, and argue that a complicated model may not give an optimal result in this respect. It is not difficult to see why in the present case. As an extreme example, we can imagine a model that counts every single presentation of a target as a different stimulus (this is in a sense true, for no two events are exactly the same. There is always a difference, say, in the lighting conditions etc.). Although such a "highly-detailed" model is guaranteed to have the highest likelihood, it says nothing about what will happen at the next presentation of a target, which it considers to be unlike any other in the past. Hence a model that best accommodates the past is not necessarily the one that serves best for predicting the future. Such a model is said to *overfit* the existing data, at the expense of its ability to predict novel data.

Estimating the predictive accuracy of a model is the principal goal of *model selection*, whose philosophical implications have been discussed by Forster and Sober [3] along with the related works [6, 13, 14]. Here, I summarize the idea. Above we saw that a model's ability to accommodate data is measured by its likelihood, the probability of observed data given that model. In contrast, the predictive ability of a model is measured by *expected likelihood*, $E(\text{data}|M)$, the likelihood averaged over all possible datasets including unobserved ones [3, 14]. The higher this value is, the better a model predicts future datasets on average. Because it concerns future and not-yet-observed data, the predictive accuracy (expected likelihood) of a model cannot be calculated from observed data, but must be estimated. Akaike [1] showed that under certain conditions, which do not concern us here, its unbiased estimator

Table 3.1 Specification of two simulation experiments. In Experiment 1, the pecking probabilities depend only on colors (RED/BLUE) and individual random effects. In Experiment 2, they also depend on difference in shades. In each experiment, the total of 10 targets are presented to each of 5 pigeons. After each experiment, Simplicio's and Complicatio's models were fitted to data and their AIC was calculated using glmmML function in R software

Color	Shade	Experiment 1	Experiment 2
Red	Dark red	0.8	0.9
	Scarlet	0.8	0.8
	Rose	0.8	0.7
Blue	Cyan	0.2	0.3
	Cobalt	0.2	0.2
	Navy	0.2	0.1

is given by

$$\log P(\mathrm{data}|M) - k$$

where k is the number of the free parameters of a model. I follow Forster and Sober [3] and call this estimate *AIC score* of model M.[2]

Akaike's results identify two factors that affect a model's predictive performance, its log-likelihood and the number of parameters. These factors often conflict: as we have seen, complex models with more parameters tend to have a higher log-likelihood, while their complexity is penalized through the second component k. Taken together, Akaike's theory tells that a model that achieves the best balance between its ability to accommodate a given dataset and simplicity will have the best average predictive accuracy.

Akaike's theory has an important implication to our discussion on proportionality. Recall that in our experimental setup, the number of parameters corresponds to a model's descriptive level: Simplicio's model has only two parameters, whereas Complicatio's has six. The question of their comparative performance thus boils down to whether the extra details/parameters introduced by Complicatio to distinguish different shades actually "pays off," i.e., boosts the log-likelihood more than the margin of 4.

The answer to this question is contingent upon the nature of the data, and to illustrate this I performed two experimental simulations under different setups. In the first simulation, pigeons are assumed to peck any reddish target at a constant probability, as shown in the third column of Table 3.1. 1000 datasets were generated from these parameters, and at each round the difference in AIC between Simplicio's

[2] This definition differs slightly from the convention in the model selection literature, where the AIC score is defined as the expected log-likelihood times negative two, i.e., $-2\log P(\mathrm{data}|M) + 2k$.

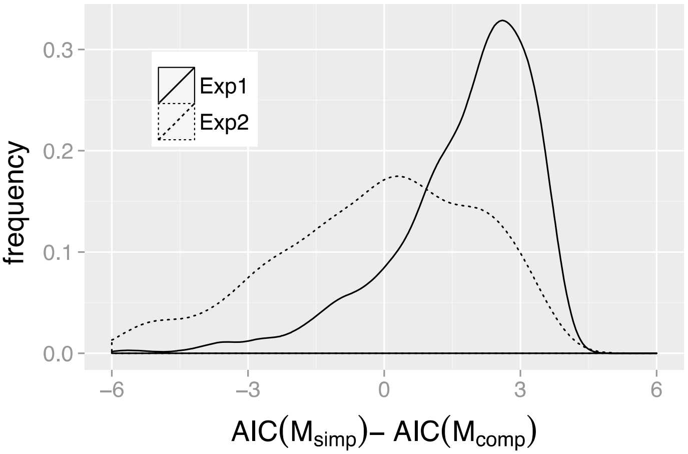

Fig. 3.1 Differences in AIC between Complicatio's and Simplicio's models, calculated from 1000 data generated each with the parameter sets in Table 3.1. In Experiment 1 (solid line), the AIC of Simplicio's model is smaller than that of Complicatio's in most cases, with the mean difference of 3.47. In contrast, the plot for Experiment 2 (dotted line) is about symmetric around zero (mean $= -0.74$)

model (M_{simp}) and Complicatio's model (M_{comp}) was calculated. The solid curve in Fig. 3.1 represents the relative frequencies of the differences under this setup, and shows that in most cases the simpler model M_{simp} scored a higher AIC. Hence in this case the AIC favors the simpler model in accordance with our intuition.

This is contrasted with mutual information. When calculated from samples in the above simulation, Complicatio's variables always had a higher sample mutual information than Simplicio's, with the means being 0.90 and 0.67, respectively (mean difference $= 0.23$ with standard deviation $= 0.10$). Hence Prop_{INF} as proposed by Pocheville et al. [10] ends up with favoring the too-detailed model in all runs despite the fact that it has no extra information. This apparent puzzle stems from random fluctuation in data. The two models will have the same mutual information only if there is no difference in the *actual* pecking rate among different shades. But the stochastic nature of the experiment means there are always slight differences, which are then counted as "extra information" in calculating sample mutual information.

Next, suppose the pigeons do differentiate shades, with the true pecking rates as shown in the rightmost column of Table 3.1 ("Experiment 2"). The dotted curve in Fig. 3.1 is the plot of $\text{AIC}\big(M_{simp}\big) - \text{AIC}\big(M_{comp}\big)$ obtained under this new setup. This time the difference in AIC between the two competitive models is less noticeable, with the mean close to zero (-0.37). This means that even though Simplicio's model is wrong, it is almost on a par in its predictive ability with Complicatio's model which better captures the reality. Truth, therefore, is not the only arbiter of models' predictive ability, but simplicity also matters; sometimes a coarse-grained model that ignores the detail of nature may be useful in predicting the future.

Elliott Sober [13, 14] has argued that Akaike's theory gives theoretical support for the use of Ockham's razor, i.e., our preference for simpler models. A similar line of argument can be made with respect to proportionality. The basic idea is that proportional variables should be preferred because they are conductive to better predictive performance. Too detailed variables, as those adopted by Complicatio, tend to require more parameters, at the cost of impairing the model's average predictive ability. On the other hand, a model must have enough granularity to correctly describe the causal relationship in question. This observation motivates us to use the AIC score to calibrate the level of description:

Prop$_{\text{AIC}}$: when comparing models M_i, $i = 0, 1, \cdots$ of different granularities, the proportionality of a model M_i with respect to data D is estimated by its AIC score, $\log P(D|M_i) - k$.

A model proportional in this sense is preferred because it is conductive to accurate predictions.

Like the previous accounts including Woodward's original definition, Prop$_{\text{AIC}}$ requires a cause to convey both enough information about its effect *and* no more than necessary. The first component, log-likelihood, measures the informativeness of the model or how well its putative cause explains the observed outcomes. The second part of the AIC, in contrast, guards against overdetailing by imposing a cost for the number of its parameters. Hence as in the original version, Prop$_{\text{AIC}}$ seeks proportionality as the balance between these two desiderata, informativeness and parsimony.

There are also dissimilarities, however. The first point of difference concerns epistemic goals. Woodward's motivation for proportionality is to obtain a simpler account of the true causal relationship, or in other words, a parsimonious picture of the reality. In contrast, Prop$_{\text{AIC}}$ is specialized for prediction tasks, favoring a simpler relationship for the sake of predictive accuracy. These two aims can conflict—the true model may not necessarily give accurate predictions, as suggested above in Experiment 2 where the predictive performance of Complicatio's true model was only little better than that of the less faithful Simplicio's model. Should the differences in parameter among shades be less significant, Simplicio's model could well have a higher AIC score. This reflects the instrumentalist character of Akaike's theory which places priority on predictive accuracy over a true description of the reality [13].

The second conspicuous difference is the explicit mention of data. The previous treatments of proportionality have questioned only the nature of functional relationships connecting causes and effects, without regard to the data with which these relationships are estimated. In contrast, Prop$_{\text{AIC}}$ explicitly depends on the data at hand, so that a model judged as proportional by one set of data may be judged otherwise by a different set. The appropriate level of description depends on how much data we have. This again comes from the nature of AIC as an estimate of predictive ability and the fact that the best predicting strategy hinges on the size of available datasets.

The next section further discusses these two characteristics in view of deriving their implications for the reductionism debate.

3.4 Multiple Realizability and Reductionism

In the philosophical literature, levels of explanation have been discussed in relation to the multiple-realizability argument against reductionism. A property A is said to multiply realize another property B if a change in the latter entails that of the former but not vice versa. In the variable notation used here, multiple realizability means that the function that maps the values of a lower-level variable to the corresponding values of a higher-level variable is non-invertible [19].[3] The existence of such a "coarse-graining" function guarantees that any state of a micro variable corresponds to a unique state of a macro variable, but not the other way around: there is at least one macro state which is multiply realized by two or more micro states. In the above pigeon experiments, Complicatio's variable describing shades multiply realizes Simplicio's color variable in this sense.

Multiple realizability has been philosophers' pet argument against reductionism. Fodor [2] claimed that because psychological states are expected to be multiply realized by a number of distinct neurological or physical states that share no non-trivial common properties, psychological generalizations can not be represented in any way but by a messy disjunction of neurological laws. Similarly, the gist of Putnam [11] famous peg-and-hole example was that the multiple realizability of the structural features of the peg and hole at the particular level makes the lower-level explanations based on the latter less general and thus inferior. These anti-reductionist arguments, however, did not go unchallenged. Sober [12] questions Fodor's premise that a disjunction of laws is not itself a law or explanatory, for many paradigmatic laws, such as "water at surface pressure will boil when it exceeds 100 °C," seem well to be disjunctive, saying that water boils at 100 °C, 101 °C, 102 °C, and so on. He also criticizes Putnam, claiming that universality is not the only desideratum of scientific explanations; one may well be interested in depth as well as breadth, and those who seek for deep explanations may legitimately prefer lower-level descriptions.

Few philosophers today doubt the explanatory relevance of higher-level sciences such as psychology or sociology. Anti-reductionists like Putnam and Fodor, however, make a stronger claim that these higher-level explanations are epistemologically *better* than lower-level counterparts, and that is in contention here. Why should we prefer macroscopic explanations? An answer suggested by the present thesis is because it provides more accurate predictions. The experiments we saw in the previous section fit Fodor's scheme of reduction, where Complicatio's predictor variable (i.e., the antecedent of his causal law) multiply realizes that of Simplicio's. As a result, Complicatio had to devise six distinct laws to express the same relationship that took Simplicio only two. The extra complexity has bought Complicatio's model a flexibility to accommodate the obtained experimental results, but did not help him to predict future outcomes. A moral here for the anti-reductionism debate is

[3] More formally: a random variable is a real-valued function defined on algebra $\mathcal{F}$ of a sample space. A random variable X *supervenes* on another Y iff for any $a, b \in \mathcal{F}$ if $X(a) \neq X(b)$ then $Y(a) \neq Y(b)$. Y *multiply realizes* X iff X supervenes on Y but not vice versa. It is easy to see that in the latter case a coarse-graining function that assigns $X(a)$ to $Y(a)$ is non-invertible.

that multiple realization and the resulting disjunctive laws of a lower-level science may lead to overfitting, which is why higher-level explanations should be (at least sometimes) preferred.

Complicatio's reductive model is said to overfit data because his variable wrongly assumes differences in causal properties where there is none or only little. In this sense, his variable does not curve the nature at its joints. But in reality the "joints" may not be so conspicuous or even discrete. In Experiment 1 one can easily recognize two causally distinct properties, Red and Blue. The distinctions among the shades in Experiment 2 are less obvious. These are just putative examples, and reality can be more subtle, with difference of order of one hundredth or one thousandth. Do these differences still mark joints? Reductionists will say yes, because ignoring such niceties, however small they are, yields a *bias* in prediction. The reductionist preference of micro variables is thus motivated and justified by the search for unbiased laws that have as few as possible exceptions.

However, the avoidance of bias in pursuit of exceptionless generalizations is not the only, nor even a major, goal of science. Another important goal is to reduce the *variance* of estimators—that is, we wish to estimate the parameters of our laws in a more precise fashion. In general, the variance gets inflated when a model contains a large number of parameters compared to the size of samples used for its estimation. Hence there is a trade-off between a model's bias and variance: the more parameters we introduce to guard our model against potential biases (and thereby making our law more disjunctive), the bigger the variance of our estimators become, and vice versa. Traditional reductionists can be seen as attaching heavy weight to the bias part of this trade-off, whereas anti-reductionists stress the variance side. But as far as predictive ability is concerned, the virtue is in the middle: Akaike's theory implies that, if the goal of finding a lawful relationship is to use it for future prediction, the best granularity must balance these two desiderata.

The key in the above discussion is the number of parameters; multiple realization impairs the predictive performance of the reducing theory provided its disjunctive laws require separate parameters. However, there are cases where multiple realization is not accompanied by an increased number of parameters, but rather enables a formulation of an even simpler law at the lower-level. Let us illustrate such a case of successful reduction with the second experiment in the previous section where the pigeons' pecking rate varied among shades (Experiment 2 of Table 3.1). Imagine that these pigeons are actually responding to light frequency so that their pecking rate is a function of frequencies of light reflected on targets. Suppose further that the frequencies of the shades are 420, 450, 480, 600, 630, and 660 THz for Dark Red, Scarlet, Rose, Cyan, Cobalt, and Navy, respectively. Now a third experimenter, Salviati, intuited this and built the following model where the pecking rate of pigeons is a liner function of light frequency X:

$$\text{Probability of pecking} = f(\alpha + \beta X), \quad \text{for some function } f \text{ and parameters } \alpha, \beta. \tag{3.1}$$

Although Salviati's X variable takes real values and is definitely finer-grained than that of the other two experimenters, his model has only two parameters, α and β. If this model is fitted to the same data used in Experiment 2, we see Salviati's model enjoys much better AIC scores than the other two models (Fig. 3.2), which suggests that Salviati's model is more accurate despite the fact that his "law" is much more disjunctive, summarizing infinite laws for each value of the real-valued variable X.

There are two reasons for the success of Salviati's model. First is the metric assumption that colors and shades come in degree and can be expressed by a ratio scale (frequencies). The metric assumption allows one not only to order color stimuli, but also to apply various arithmetic operations such as addition or multiplication. This insight presumably comes from knowledge of optic theory, and provides a deeper understanding of the nature of the cause variable X. The second key factor for the success is the functional assumption that the shades thus expressed are systematically related to the pecking rate via Eq. (3.1). This formula assumedly summarizes a theory about the complex neurological and physiological mechanisms relating visual stimuli to pigeons' behavior. Salviati's model thus stands on the shoulders of these elaborated theories, which make his law distinct from mere disjunctions. The difference is a systematic relationship—Salviati's law (3.1) does not just tell us the pecking rate for each target, but does so *systematically*. This is also the reason why the law about the boiling point of water mentioned by Sober [12] should be distinguished from what Fodor [2] had in mind when he dismissed disjunctive laws as non-explanatory. Temperature is measured by interval scale, and already has a rich metric structure to it; hence, saying that water boils when it exceeds 100 °C is *not* the same as saying that it boils at 100 °C, 101 °C, and so on.

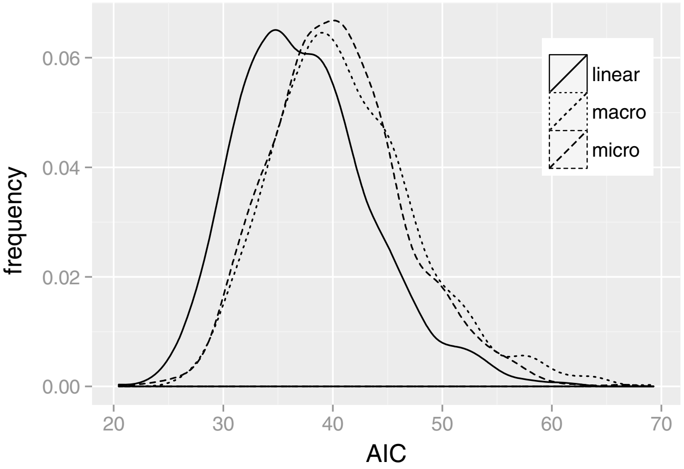

Fig. 3.2 Comparison of AIC among Linear, Macro, and Micro models. Even though Salviati's X variable is much finer-grained than that of the other two, his linear model scores smaller AICs (solid line) and is expected to provide more accurate predictions

A successful reduction happens, therefore, when a scientific theory enables us to formulate a systematic relationship at the lower level in a simple way. Of course such a theory and relationships are hard to come by in most special sciences such as biology, psychology, sociology, and so on; the only reason Salviati could come up with his nice solution above is that I made it so. In reality there are objective and epistemological challenges. First, it may simply be the case that nature at its microscopic scale lacks systematic relationships, as Fodor [2] surmised. Or even if they exist, these relationships may forever stay hidden from our scientific investigations.

In addition to these two obstacles for successful reduction, the model selection perspective suggests a third, pragmatic factor that should be considered in the discussion of reductionism. The pragmatic consideration comes from the nature of AIC as a tool for evaluating the average predictive accuracy of a model [14]. Which model is considered the best tool naturally depends on our goal, or the size of data used to fit a model. For example, a model suited for predicting small datasets does not necessarily fare well with large datasets. We have seen that Simplicio's and Complicatio's models almost tied in Experiment 2; but with a bigger sample size (e.g., with 10 instead of 3 targets presentations for each trial of 10 instead of 5 pigeons), Complicatio's model outcompeted Simplicio's with the mean difference in their respective AIC scores $\mathrm{AIC}\big(M_{comp}\big) - \mathrm{AIC}\big(M_{simp}\big) = 11.4$. Thus under this data-rich situation $\mathrm{Prop}_{\mathrm{AIC}}$ favors Complicatio's reductive model as being more proportional. In general, increasing sample size allows for finer-tuning of reductive models. An appropriate level of description, therefore, depends on the size of data at our disposal. If we have a large dataset it makes more sense to adopt fine-grained models, but otherwise we might be better staying at a macro level.

3.5 Objective, Epistemological, and Pragmatic Aspects of Reduction

The relationship among the three—objective, epistemological, and pragmatic—aspects of reduction mentioned above can further be clarified by comparing AIC and mutual information. Above we saw mutual information $I(X;Y)$ as a measure of the amount of information X carries with respect to Y. There is an alternative interpretation, based on the following identity

$$I(X;Y) = \mathrm{KL}(P(X,Y);\,P(X)P(Y)), \tag{3.2}$$

where KL is the *Kullback-Leibler divergence* (KL divergence), an information theoretic measure of the distance between two distributions.[4] From this perspective mutual information measures the distance of the product of marginal distribution $P(X)P(Y)$ from the joint distribution $P(X,Y)$. If this distance is zero then

[4] To be precise KL divergence is not a distance because it is not symmetric—i.e., $\mathrm{KL}(f,g) \neq \mathrm{KL}(g,f)$ in general. This, however, is not relevant to the discussion here.

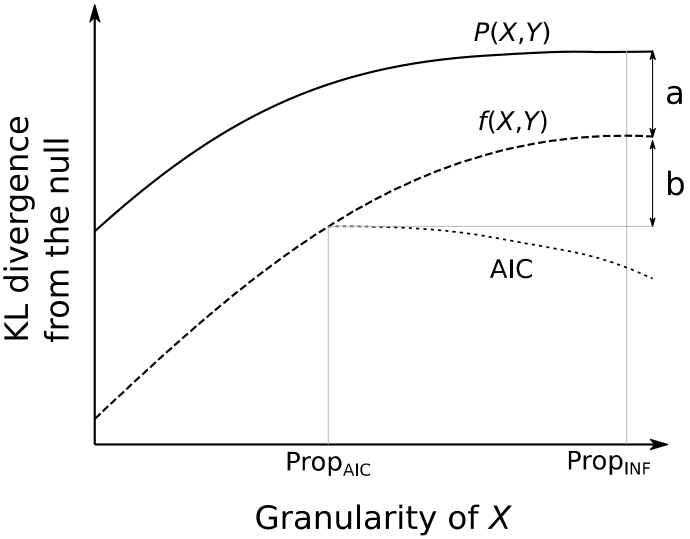

Fig. 3.3 A hypothetical plot of the (estimated) distance of various indices from the null model $P(X)P(Y)$ for different granularities of X. **Solid curve**: the true joint distribution $P(X, Y)$ sets the upper bound of the information one can exploit by relating X and Y. **Dashed curve**: as X becomes detailed a model $f(X, Y)$ approaches the truth $P(X, Y)$, but does not reach it. The remaining distance **a** is due to our ignorance of the true distribution. **Dotted curve**: the expected distance from the true distribution of a model fitted with finite sample may increase if a finer description introduces more parameters. **b** Represents the loss due to a pragmatic constraint on available sample size

$P(X, Y) = P(X)P(Y)$, namely X and Y are independent and thus there is no point in considering X and Y together in the form of joint distribution. In contrast, a large distance suggests that treating X and Y separately likely misses the whole picture. Mutual information thus measures the modeling opportunity—that is, how worthwhile it is to relate X to Y to begin with.

Now, consider plotting $I(X; Y)$ for different granularities of X variable (solid curve in Fig. 3.3). The horizontal axis of the plot represents granularity of X, where a variable at each point multiply realizes all the variables to the left.[5] The vertical axis measures, for a given level of X, how much information it has about Y, or equivalently from (3.2), the distance of the joint distribution from the "null-model" $P(X)P(Y)$ where X and Y are treated unrelatedly. Since detailing a variable never gets rid of the information it already has, the solid curve is non-decreasing, but the steepness of the slope depends on the nature of the relationship between X and Y. If it is steep, we can exploit more information about the effect by further detailing the cause, i.e., there is a lot of opportunity for reduction. In contrast, a flat slope means that higher-level properties already exhaust most of the potential information about the effect. The slope of $I(X; Y)$, therefore, reflects the objective constraint

[5] Since multiple realization forms a partial order, there are multiple, possibly infinite, ways to align variables according to their granularity. The X-axis of the plot is just one of them.

on reduction imposed by the nature of the causal relationship connecting the two variables. The proposal of Pocheville et al. [10] is to find the plateau of this curve, i.e. the coarsest X that can exhaust all the information of Y we can get by knowing their true relationship.

In contrast to the objective constraint that pertains to the nature of the relationship and is encoded in the true joint distribution $P(X, Y)$, the epistemological constraint on reduction stems from our ignorance. For want of the true picture, we build a model $f(X, Y)$ that we think approximates $P(X, Y)$. Since a model is only an approximation of the truth, it has a nonzero KL divergence from the true distribution and is positioned somewhere between $P(X, Y)$ and the "null-model" $P(X)P(Y)$ in Fig. 3.3. This KL divergence is negatively proportional to the expected log likelihood of the model, the value that AIC tries to estimate.

$$\mathrm{KL}(P(X, Y); f(X, Y)) = \mathrm{Const.} - E\big[\log f(X, Y)\big]. \tag{3.3}$$

In reality the expected log likelihood stays unknown (the right-hand side is an expectation over the true probability distribution) and thus must be estimated from finite samples, say via AIC. But here let us assume our limitation is only epistemic, and we have infinite data to correctly determine the expected log likelihood of the model with different granularities of X. Under this assumption, a model's expected log likelihood never decreases as its variable gets fine-grained, which means the KL divergence of the model from the true distribution is non-increasing (dashed curve in Fig. 3.3). The actual slope of the curve depends on the model. A model showing a steep slope, for example, will approximate well the truth on microscopic scales, and thus has a high potential for reduction. The KL divergence (3.3) of the model from the true distribution thus represents the loss of the opportunity of reduction due to the epistemic limitation of not knowing the true distribution.

Finally, where does the pragmatic factor fit in this plot? The pragmatic limitation relevant to the current discussion concerns our data-gathering ability. With a finite sample, a model's predictive performance depends on whether we have enough data to afford its complexity. AIC is formally derived as an estimate of the average KL divergence of the distribution predicted by a model from data sampled from the true distribution. This distance may *increase* as a model gets finer-grained, as we have seen in our simulation experiments. Maximizing the AIC score among models with different granularities amounts to minimizing the distance between the solid curve and the dotted curve in Fig. 3.3. The best or most proportional model in this sense, indicated by $\mathrm{Prop}_{\mathrm{AIC}}$ in the figure, tends to be coarser compared to the optimal model under infinite sample size, with the difference between them (b in Fig. 3.3) representing the pragmatic limitation on our data-gathering ability.

The objective, epistemological, and pragmatic limitations for reduction can thus be visualized as divergences from the null model. Note that this plot is just for illustration and not meant to be a representative case; the shape of the curves depends on the nature of the relationship and model in question. Qualitative remarks about the figure, however, are general: (i) The information X contains about *Y with regard to the true distribution* is never lost as X gets finer-grained, thus the solid curve is always non-decreasing. (ii) The information X contains about Y *with regard to a model* (dashed curve) is also non-decreasing, but does not reach the mutual information. (iii) A model's actual performance as estimated by AIC with finite samples (dotted curve) *may decrease* as X gets fine-grained. (iv) For these reasons the AIC-based proportionality (Prop_{AIC}) is always coarser than that based on mutual information (Prop_{INF}). The loss of information (a + b) due to this coarse-graining reflects the two limitations discussed above, namely our ignorance of the true distribution and limited data to fit a hypothesized model.

The plot also helps us to understand various attitudes toward reductionism. First, one may construe the problem of reduction as an in-principle matter that concerns the true picture of the world or ideally completed sciences. The primary question on this construal would be which level faithfully captures all there is to know about the relationship between two variables, or maximizes their mutual information. If this is the problem, reduction to a lower-level science "never hurts," for mutual information (solid curve) is a non-decreasing function of granularity. There may be a point, Prop_{INF}, beyond which no further reduction yields additional information and thus is *unnecessary*, but nevertheless *innocuous*. The objective, in-principle attitude thus admits only this kind of weak form of anti-reductionist stopping rule.

Next, those who take the inherent incompleteness of our scientific knowledge seriously might be interested in the epistemological merit of reduction, and would ask whether reduction improves our theory by bringing it closer to the truth. Their question, then, is which level minimizes the KL divergence of a model from the true distribution, that is, the distance between the solid and dashed curves in Fig. 3.3. The shift in question, however, does not affect the overall inclination toward reductionism. Because the KL divergence in question is non-increasing, there is no penalty for a lower-level variable; any model is at least as close to the truth as its coarse-grained version that uses a multiply realized variable. Hence this construal too motivates only the weak anti-reductionism.

Finally, consider a more realistic stance that acknowledges not only the incompleteness of scientific knowledge but also the limit of our data-gathering ability. The question then is which level best serves our epistemic purposes given finite data available in a specific research context. In this case reduction is not always good, at least for the purpose of predictions; reducing variables beyond a certain granularity marked by Prop_{AIC} is not just otiose but potentially harmful for a model's predictive performance (dotted curve). Hence the focus on the pragmatic limitation motivates the *strong form of anti-reductionism* that cautions against a definite demerit of reductive investigation.

Woodward's functional account best fits with the last among these three attitudes towards reductionism, with its focus on "usefulness of different causal concepts, and of procedures for relating causal claims to evidence" [18], p. 694). Evaluating usefulness makes sense only in relation to their users who are limited in both knowledge and resource. The advantage of the AIC-based approach presented here is its explicit recognition of these limitations from which it derives the positive value of proportionality, namely, that proportional variables can be more useful despite containing less information than other finer-grained descriptions.

An implication of this is that proportionality is a pragmatic standard, rather than an epistemic criterion for truth. Although some philosophers expressed a concern that the focus on pragmatics introduces some kind of anthropocentrism into scientific explanations [4], I argue that it is a virtue rather than a vice of our account. First of all, very few, if any, philosophers today would deny the pragmatic dimensions of scientific practices and explanations [16]. Moreover, a consideration of pragmatic factors proves essential in the context of sociological studies, policy making, and risk analyses, where the range and amount of possible observations are severely limited by practical, financial, and ethical reasons. Facing complex social issues, scientists and policy makers must limit their research focus on only a tiny fragment of possibly relevant factors and draw a conclusion based on a relatively small dataset. In such situations, it makes much more sense to let your model and conclusion depend on pragmatic factors rather than on some epistemic ideal one can never attain. In this respect, the pragmatic aspect of the present approach is not a philosophical drawback, but rather a necessary element to understand our explanatory practices.

3.6 Conclusion

Justifying the use of high-level explanations in the so-called special sciences has long been a major challenge in the reductionism debate and in philosophical theories of explanation in general. The proportionality criterion was proposed to save high-level causal explanations, but its precise formulation and, more importantly, its epistemological merit have come under discussion. This paper offered a new interpretation of proportionality based on the Akaike Information Criterion. AIC-based proportionality estimates the predictive accuracy of a model by balancing its bias and variance. A model with a too detailed variable tends to overfit data, which impairs its predictive performance. In such cases we should prefer macroscopic and less detailed explanations over microscopic ones.

The chapter illustrated this with a rather simplistic example of pigeons, but one can easily imagine a similar explanatory task arises in social sciences. For instance, a researcher may be interested in whether the political climate of a country affects its ratification of an international pact on some cause, say environmental protection. The explanatory variable here can be described at various levels: one can, like Simplicio, dichotomize it into either conservative or liberal, or, as did Complicatio, adopt a finer sub-categorization of political spectrum that distinguishes neo-liberalism, social democracy, green party, populism, etc. The later by definition gives a more detailed picture, but not necessarily a better prediction as to whether a new country ratifies the pact in question. As we have seen, which descriptive level the researcher should choose depends on data available to fit the models as well as the nature of the problem.

This conclusion sheds new light on the long-standing debate on reductionism in social sciences. The discussion between methodological individualism and holism has mainly resolved around the in-principle derivability of macro states, properties or theories from micro counterparts [9]. But as Lohse argues, the choice between individualist versus holist explanation should depend "on our epistemic interest (what do we want to know?) and pragmatic aspects such as efficiency [8]." The AIC-based approach takes into account this pragmatic aspect by evaluating the "efficiency" of explanations/models of different granularity in terms of their predictive ability. Since the AIC score depends on objective, epistemic, and pragmatic factors which are all case-relative, the present approach supports the local, piece-meal view of reduction rather than the classical view that focuses on the derivability of entire theories [12, 17]. According to the piece-meal view, whether we should adopt a reductive or "individualist" explanation or not should be determined not by an in-principle fiat, but rather by case-by-case considerations on empirical as well as pragmatic circumstances of the problem at hand. The model selection perspective described in this paper clarifies which factors should be accounted for in each of such decisions, and why.

The focus on pragmatics is in line with the "functional approach" [18], which takes usefulness in actual scientific practices as an important (or in Woodward's view, the only) arbiter of philosophical accounts of explanations. Usefulness in the present context meant predictive accuracy. This particular choice reflects our use of causal models to predict (intervention) consequences, but I by no means claim this to be the only criterion of usefulness. As another research context not so much related to prediction, one may be interested in finding a descriptive level of longitudinal data in which any variable of the auto-regression model screens off all the prior variables from subsequent ones. The AIC-based proportionality as proposed in this paper falls short for such a purpose because it does not guarantee the desired Markov property. The current proposal, therefore, should be understood as *an* interpretation of proportionality for the purpose of prediction. How do different epistemic goals affect our choice of description remains to be seen.

References

1. Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.
2. Fodor, J. A. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese, 28*(2), 97–115.
3. Forster, M., & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *British Journal for the Philosophy of Science, 45*(1), 1–35.
4. Franklin-Hall, L. R. (2016). High-level explanation and the interventionist's 'variables problem'. *The British Journal for the Philosophy of Science, 67*(2), 553–577.
5. Griffiths, P. E., Pocheville, A., Calcott, B., Stotz, K., Kim, H., & Knight, R. (2015). Measuring causal specificity. *Philosophy of Science, 82*(4), 529–555.
6. Hitchcock, C., & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *The British Journal for the Philosophy of Science, 55*(1), 1–34.
7. Kincaid, H. (2017). Reductionism. In L. McIntyre & A. Rosenberg (Eds.), *The Routledge companion to philosophy of social science* (pp. 113–123), New York: Routledge.

8. Lohse, S. (2016). Pragmatism, ontology, and philosophy of the social sciences in practice. *Philosophy of The Social Sciences, 47*(1), 3–27.
9. McGinley, W. (2011). Reduction in sociology. *Philosophy of the Social Sciences, 42*(3), 370–398.
10. Pocheville, A., Griffiths, P. E., & Stotz, K. (2016). Comparing causes: An information-theoretic approach to specificity, proportionality and stability. In H. Leitgeb, I. Niiniluoto, E. Sober, & P. Seppälä editors (Eds.), *Proceedings of the 15th Congress of Logic, Methodology and Philosophy of Science.*
11. Putnam, H. (1975). Philosophy and our mental life. In *Mind, language, and reality* (pp. 291–303). Cambridge: Cambridge University Press.
12. Sober, E. (1999). The multiple realizability argument against reductionism. *Philosophy of Science, 66,* 542–564.
13. Sober, E. (2002). Instrumentalism, parsimony, and the akaike framework. *Philosophy of Science, 69*(S3), S112–S123.
14. Sober, E. (2008). *Evidence and evolution.* Cambridge: Cambridge University Press.
15. Sober, E. (2015). *Ockham's Razors. A user's manual.* Cambridge: Cambridge University Press.
16. van Fraassen, B. C. (1980). *The scientific image.* Oxford: Oxford University Press.
17. Woodward, J. (2010). Causation in biology: Stability, specificity, and the choice of levels of explanation. *Biology and Philosophy, 25*(3), 287–318.
18. Woodward, J. (2014). A functional account of causation: Or, a defense of the legitimacy of causal thinking by reference to the only standard that matters—usefulness (as opposed to metaphysics or agreement with intuitive judgment). *Philosophy of Science, 81*(5), 691–713.
19. Woodward, J. (2016). The problem of variable choice. *Synthese, 193*(4), 1047–1072.
20. Yablo, S. (1992). Mental Causation. *The Philosophical Review, 101*(2), 245–280.
21. Zahle, J. (2016). The individualism-holism debate on intertheoretic reduction and the argument from multiple realization. *Philosophy of the Social Sciences, 33*(1), 77–99.