ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2017)

Beyond subjective and objective in statistics

Andrew Gelman

Columbia University, New York, USA

and Christian Hennig

University College London, UK

[*Read before* The Royal Statistical Society *on Wednesday, April 12th, 2017,* Professor P. J. Diggle *in the Chair*]

Summary. Decisions in statistical data analysis are often justified, criticized or avoided by using concepts of objectivity and subjectivity. We argue that the words 'objective' and 'subjective' in statistics discourse are used in a mostly unhelpful way, and we propose to replace each of them with broader collections of attributes, with objectivity replaced by *transparency*, *consensus*, *impartiality* and *correspondence to observable reality*, and subjectivity replaced by awareness of *multiple perspectives* and *context dependence*. Together with *stability*, these make up a collection of virtues that we think is helpful in discussions of statistical foundations and practice. The advantage of these reformulations is that the replacement terms do not oppose each other and that they give more specific guidance about what statistical science strives to achieve. Instead of debating over whether a given statistical method is subjective or objective (or normatively debating the relative merits of subjectivity and objectivity in statistical practice), we can recognize desirable attributes such as transparency and acknowledgement of multiple perspectives as complementary goals. We demonstrate the implications of our proposal with recent applied examples from pharmacology, election polling and socio-economic stratification. The aim of the paper is to push users and developers of statistical methods towards more effective use of diverse sources of information and more open acknowledgement of assumptions and goals.

Keywords: Bayesian; Frequentist; Good practice; Philosophy of statistics; Virtues

1. Introduction

We cannot do statistics without data, and as statisticians much of our efforts revolve around modelling the links between data and substantive constructs of interest. We might analyse national survey data on purchasing decisions as a way of estimating consumers' responses to economic conditions, or gather blood samples over time on a sample of patients with the goal of estimating the metabolism of a drug, with the ultimate goal of coming up with a more effective dosing schedule; or we might be performing a more exploratory analysis, seeking clusters in a multivariate data set with the aim of discovering patterns that are not apparent in simple averages of raw data.

As applied researchers we are continually reminded of the value of integrating new data into an analysis, and the balance between data quality and quantity. In some settings it is possible to answer questions of interest by using a single clean data set, but increasingly we are finding that this simple textbook approach does not work.

Address for correspondence: Andrew Gelman, Department of Statistics, Columbia University, Room 1016, 1255 Amsterdam Avenue, New York, NY 10027, USA.
E-mail: gelman@stat.columbia.edu

 0964–1998/17/180000

External information can come in many forms, including

(a) recommendations on what variables to adjust for non-representativeness of a survey or imbalance in an experiment or observational study,
(b) the extent to which outliers should be treated as regular, or erroneous, or as indicating something that is meaningful but essentially different from the main body of observations,
(c) issues of measurement, confounding, and substantively meaningful effect sizes,
(d) population distributions that are used in poststratification, age adjustment and other procedures that attempt to align inferences to a common population of interest,
(e) restrictions such as smoothness or sparsity that serve to regularize estimates in high dimensional settings,
(f) the choice of the functional form in a regression model (which in economics might be chosen to work with a particular utility function, or in public health might be motivated on the basis of success in similar studies in the literature) and
(g) numerical information about particular parameters in a model.

Of all these, only the final item is traditionally given the name 'prior information' in a statistical analysis, but all can be useful in serious applied work. Other relevant information concerns not the data-generating process but rather how the data and results of an analysis are to be used or interpreted.

We were motivated to write the present paper because we felt that our applied work, and that of others, was impeded because of the conventional framing of certain statistical analyses as subjective. It seemed to us that, rather than being in opposition, subjectivity and objectivity both had virtues that were relevant in making decisions about statistical analyses. We have earlier noted (Gelman and O'Rourke, 2015) that statisticians typically choose their procedures on the basis of non-statistical criteria, and philosophical traditions and even the labels attached to particular concepts can affect real world practice.

In Section 2 we show how the discussions of objectivity and subjectivity affect statistical practice and statistical thinking, followed by an outline of our own philosophical attitude to objectivity and subjectivity in science; Appendix A provides an overview of what the philosophy of science has to say on the matter. In Section 3 we present our proposal, exploding objectivity and subjectivity into several specific virtues to guide statistical practice. In Section 4 we demonstrate the relevance of these ideas for three of our active applied research projects: a hierarchical population model in pharmacology, a procedure for adjustment of opt-in surveys and a cluster analysis of data on socio-economic stratification. In Section 5 we revisit fundamental approaches to statistics by using the proposals of Section 3, demonstrating how they can elaborate advantages and disadvantages of the various approaches in a more helpful way than the traditional labelling of 'objective' and 'subjective'. Section 6 contains a final discussion, including a list of issues in scientific publications that could be addressed by using the virtues that are proposed here.

2. Objectivity and subjectivity

2.1. Objectivity, subjectivity and decision making in statistics

Concepts of objectivity and subjectivity are often used to justify, criticize, avoid or hide the decisions that are made in data analysis. Typically, the underlying idea is that science should be objective, understood as something like 'independence of personal biases', without referring to any clear definition. Concepts of objectivity can be implicitly invoked when making choices, so that certain decisions are avoided or hidden in order not to open an analysis to charges of subjectivity.

For many statistical methods tuning constants need to be decided such as the proportion of trimmed observations when computing a trimmed mean or bandwidths for smoothing in density estimation or non-parametric regression; one could also interpret the conventional use of the 0.05 level of significance as a kind of tuning parameter. In the statistical literature, methods are advertised by stating that they do not require any tuning decisions by the user. Often these choices are hidden so that users of statistics (particularly those without specific background knowledge in statistics) expect that for their data analysis task there is a unique correct statistical method. This expectation is exploited by the marketing strategies for some data analysis software packages that suggest that a default analysis is only one click away. Although such an approach obviously tempts the user by its simplicity, it also appeals on the level of avoiding individual impact or subjectivity. Decisions that need to be made are taken out of the hand of the user and are made by the algorithm, removing an opportunity for manipulation but ignoring valuable information about the data and their background. This is in stark contrast with the typical trial-and-error way of building one or more statistical models with plenty of subjective decisions starting from data preprocessing via data exploration and choice of method onto the selection of how to present which results. Realistically, even one-click methods require user choices on data coding and data exclusion, and these inputs can have big influences on end results such as *p*-values and confidence intervals (Steegen *et al*., 2006).

More mathematically oriented statisticians design and choose methods that optimize criteria such as unbiased minimum variance, often relying on restrictive assumptions. This can allow for elegant theoretical results with narrow scope. When methods are to be compared in simulation studies, typically there is a huge variety of choices including data-generating processes (distributions, parameters, dimensionality, etc.), performance measures and tuning of competitors. This can easily discourage researchers from running such studies at all or at least beyond one or two illustrative toy set-ups but, despite their subjective flavour and the difficulty in finding a path through a potentially confusing jungle of results, such studies can be informative and raise important issues. Over 50 years ago, Tukey criticized an obsession with mathematical optimization problems concentrating on one-dimensional criteria and simple parametric models in the name of objectivity, and stated that 'in data analysis we must look to a very heavy emphasis on judgment' (Tukey, 1962).

Researchers often rely on the seeming objectivity of the $p < 0.05$ criterion without realizing that theory behind the *p*-value is invalidated when analysis is contingent on data (Simmons *et al*., 2011; Gelman and Loken, 2014). Significance testing can be part of a misguided ideology that leads researchers to hide, even from themselves, the iterative searching process by which a scientific theory is mapped into a statistical model or choice of data analysis (Box, 1983). More generally, overreaction to concerns about subjectivity can lead researchers to avoid incorporating relevant and available information in their analyses and not to adapt analyses appropriately to their research questions and potential uses of their results.

Personal decision making cannot be avoided in statistical data analysis and, for want of approaches to justify such decisions, the pursuit of objectivity degenerates easily to a pursuit to merely *appear* objective. Scientists whose methods are branded as subjective have the awkward choice of either saying, 'No, we are really objective', or else embracing the subjective label and turning it into a principle, and the temptation is high to avoid this by hiding researcher degrees of freedom from the public unless they can be made to appear 'objective'. Such attitudes about objectivity and subjectivity can be an obstacle to good practice in data analysis and its communication, and we believe that researchers can be guided in a better way by a list of more specific scientific virtues when choosing and justifying their approaches.

The continuing interest in and discussion of objectivity and subjectivity in statistics are, we believe, a necessary product of a fundamental tension in science: on one hand, scientific claims should be impersonal in the sense that a scientific argument should be understandable by anyone with the necessary training, not just by the person promulgating it, and it should be possible for scientific claims to be evaluated and tested by outsiders. On the other hand, a reality that is assumed to be objective in the sense of being independent of its observers is only accessible through observations that are made by observers and dependent on their perspectives; communication about the observations and the process of observation and measurement relies on language constructs. Thus objective and subjective elements arise in the practice of science, and similar considerations hold in statistics.

Within statistics, though, discourse on objectivity and subjectivity is at an impasse. Ideally these concepts would be part of a consideration of the role of different sorts of information and assumptions in statistical analysis, but instead they often seemed to be used in restrictive and misleading ways.

One problem is that the terms 'objective' and 'subjective' are loaded with so many associations and are often used in a mixed descriptive–normative way. For example, a statistical method that does not require the specification of any tuning parameters is objective in a descriptive sense (it does not require decisions by the individual scientist). Often this is presented as an advantage of the method without further discussion, implying objectivity as a norm, but the lack of flexibility that is caused by the impossibility of tuning can actually be a disadvantage (and indeed can lead to subjectivity at a different point in the analysis, when the analyst must make the decision of whether to use an autotuned approach in a setting where its inferences do not appear to make sense). The frequentist interpretation of probability is objective in the sense that it locates probabilities in an objective world that exists independently of the observer, but the definition of these probabilities requires a subjective definition of a reference set. Some proponents of frequentism consider its objectivity (in the sense of impersonality, conditional on the definition of the reference set) as a virtue, but this property is ultimately only descriptive; it does not imply on its own that such probabilities indeed exist in the objective world, nor that they are a worthwhile target for scientific inquiry.

In discussions of the foundations of statistics, objectivity and subjectivity are seen as opposites. Objectivity is typically seen as a good thing; many see it as a major requirement for good science. Bayesian statistics is often presented as being subjective because of the choice of a prior distribution. Some Bayesians (notably Jaynes (2003) and Berger (2006)) have advocated an objective approach, whereas others (notably de Finetti (1974)) have embraced subjectivity. It has been argued that the subjective–objective distinction is meaningless because all statistical methods, Bayesian or otherwise, require subjective choices, but the choice of prior distribution is sometimes held to be particularly subjective because, unlike the data model, it cannot be determined even in the asymptotic limit. In practice, subjective prior distributions often have well-known empirical problems such as overconfidence (Alpert and Raiffa, 1984; Erev *et al.*, 1994), which motivates efforts to check and calibrate Bayesian models (Rubin, 1984; Little, 2012) and to situate Bayesian inference within an error statistical philosophy (Mayo, 1996; Gelman and Shalizi, 2013).

de Finetti can be credited with acknowledging honestly that subjective decisions cannot be avoided in statistics, but it is misleading to think that the required subjectivity always takes the form of prior belief. The confusion arises from two directions: first, prior distributions are not necessarily any more subjective than other aspects of a statistical model; indeed, in many applications priors can and are estimated from data frequencies (see chapter 1 of Gelman *et al.* (2013) for several examples). Second, somewhat arbitrary choices come into many aspects of

statistical models, Bayesian and otherwise, and therefore we think that it is a mistake to consider the prior distribution as the exclusive gate at which subjectivity enters a statistical procedure.

On one hand, statistics is sometimes said to be the science of defaults: most applications of statistics are performed by non-statisticians who adapt existing general methods to their particular problems, and much of the research within the field of statistics involves devising, evaluating and improving such generally applicable procedures (Gelman, 2014a). It is then seen as desirable that any required data analytic decisions or tuning are performed in an objective manner, either determined somehow from the data or justified by some kind of optimality argument.

On the other hand, practitioners must apply their subjective judgement in the choice of what method to use, what assumptions to invoke and what data to include in their analyses. Even using 'no need for tuning' as a criterion for method selection or prioritizing bias, for example, or mean-squared error, is a subjective decision. Settings that appear completely mechanical involve choice: for example, if a researcher has a checklist saying to apply linear regression for continuous data, logistic regression for binary data and Poisson regression for count data, he or she still has the option of coding a response as continuous or to use a threshold to define a binary classification. And such choices can be far from trivial; for example, when modelling elections or sports outcomes, one can simply predict the winner or instead predict the numerical point differential or vote margin. Modelling the binary outcome can be simpler to explain but in general will throw away information, and subjective judgement arises in deciding what to do in this sort of problem (Gelman, 2014b). And, in both classical and Bayesian statistics, subjective choices arise in defining the sample space and considering what information to condition on.

2.2. Objectivity, subjectivity and quantification in scientific measurement

Another issue that is connected to objectivity and subjectivity relevant to statisticians has to do with where the data to be analysed come from. There is an ideology that is widespread in many areas of science that sees quantification and numbers and their statistical analysis as key tools for objectivity. An important function of quantitative scientific measurement is the production of observations that are thought of as independent of individual points of view. But, even apart from the generally difficult issue of measurement validity, the focus on what can be quantified can narrow down what can be observed and may not necessarily do the measured entities justice.

The social sciences have seen endless arguments over the relative importance of objective conditions and what Keynes (1936) called 'animal spirits'. In macroeconomics, for example, the debate has been between the monetarists who tend to characterize recessions as necessary consequences of underlying economic conditions (as measured, for example, by current account balances, business investment and productivity), and the Keynesians who focus on more subjective factors such as stock market bubbles and firms' investment decisions. These disagreements also turn methodological, with much dispute, for example, over the virtues and defects of various attempts to measure objectively the supply and velocity of money, or consumer confidence or various other inputs to economic models. In psychology, there is a big effort to measure personality traits and subjective states scientifically. For example, Kahneman (1999) defined 'objective happiness' as 'the average of utility over a period of time'. Whether or not this definition makes much sense, it illustrates a movement in the social and behavioural sciences to measure, in supposedly objective manners, what might previously have been considered unmeasurable. Another example is the use of quantitative indicators for human rights in different countries; although it has been argued that it is of major importance that such indicators should be objective to have appropriate impact on political decision making (Candler *et al.*, 2011), many aspects of their

definition and methodology are subject to controversy and reflect specific political interests and views (Merry, 2011), and we think that it will help the debate to communicate such indicators transparently together with their limitations and the decisions involved rather than to sell them as objective and unquestionable.

Connected to quantification as a means of objectification is an attitude to statistics of many researchers in various areas who use standard routines in statistical software without much understanding of how the methods' assumptions and motivation relate to their specific research problem, in the expectation that the software can condense their research into a single summary (most often a p-value) that 'objectifies' their results. This idea of objectivity is in stark contrast with the realization by many of these researchers at some point that depending on individual inventiveness there are many ways to arrive at such a number.

See Porter (1996), Desrosieres (2002) and Douglas (2009) for more discussion of the connection between quantification and objectivity. As with choices in statistical modelling and analysis, we believe that when considering measurement the objective–subjective antagonism is less helpful than a more detailed discussion of what quantification can achieve and what its limitations are.

2.3. Our attitude towards objectivity and subjectivity in science

Many users of the terms 'objective' and 'subjective' in discussions concerning statistics do not acknowledge that these terms are quite controversial in the philosophy of science (as is 'realism') and that they are used with a variety of meanings and are therefore prone to misunderstandings. An overview is given in Appendix A.

The attitude that is taken in the present paper is based on Hennig (2010). According to this perspective, human inquiry starts from observations that are made by personal observers ('personal reality'). Through communication, people share observations and generate 'social realities' that go beyond a personal point of view. These shared realities include for example measurement procedures that standardize observations, and mathematical models that connect observations to an abstract formal system that is meant to create a thought system that is cleaned from individually different points of view. Nevertheless, human beings only have access to 'observer-independent reality' through personal observations and how these are brought together in social reality.

Science aims at arriving at a view of reality that is stable and reliable and can be agreed freely by general observers and is therefore as observer independent as possible. In this sense we see objectivity as a scientific ideal. But at the same time we acknowledge what gave rise to the criticism of objectivity: the existence of different individual perspectives and also of perspectives that differ between social systems, and therefore the ultimate inaccessibility of a reality that is truly independent of observers is a basic human condition. Objectivity can only be attributed by observers and, if observers disagree about what is objective, there is no privileged position from which this can be decided. Ideal objectivity can never be achieved.

How to resolve scientific disputes by scientific means without throwing up our hands and giving up on the possibility of scientific consensus is a key problem, and science should be guided by principles that at the same time aim at stable and reliable consensus as usually associated with 'objectivity' while remaining open to a variety of perspectives, often associated with 'subjectivity', exchange between which is needed to build a stable and reliable scientific world view.

Although there is no objective access to observer-independent reality, we acknowledge that there is an almost universal human experience of a reality perceived as located outside the

observer and as not controllable by the observer. We see this reality as a target of science, which makes observed reality a main guiding light for science. We are therefore 'active scientific realists' in the sense of Chang (2012), who wrote

> 'I take reality as whatever is not subject to one's will, and knowledge as an ability to act without being frustrated by resistance from reality. This perspective allows an optimistic rendition of the pessimistic induction, which celebrates the fact that we can be successful in science without even knowing the truth. The standard realist argument from success to truth is shown to be ill-defined and flawed.'

Or, more informally, 'Reality is that which, when you stop believing in it, doesn't go away' (Dick, 1981). Active scientific realism implies that finding out the truth about objective reality is not the ultimate aim of science, but that science rather aims at supporting human actions. This means that scientific methodology must be assessed relatively to the specific aims and actions that are connected to its use.

Because science aims at agreement, communication is central to science, as are transparency and techniques for supporting the clarity of communication. Among these techniques are formal and mathematical language, standardized measurement procedures and scientific models. Such techniques provide a basis for scientific discussion and consensus, but at the same time the scientific consensus should not be based on authority and it always needs to be open to new points of view that challenge an established consensus. Therefore, in science there is always a tension between the ideal of general agreement and the reality of heterogeneous perspectives, and the virtues that are listed in Section 3 are meant to help statisticians navigating this tension.

3. Our proposal

To move the conversation towards principles of good science, we propose to replace, wherever possible, the words 'objectivity' and 'subjectivity' with broader collections of attributes, namely by *transparency*, *consensus*, *impartiality* and *correspondence to observable reality*, all related to objectivity, awareness of *multiple perspectives* and *context dependence*, related to subjectivity, and *investigation of stability*, related to both.

The advantage of this reformulation is that the replacement terms do not oppose each other. Instead of debating over whether a given statistical method is subjective or objective (or normatively debating the relative merits of subjectivity and objectivity in statistical practice), we can recognize attributes such as transparency and acknowledgement of multiple perspectives as complementary.

3.1. 'Transparency', 'consensus', 'impartiality' and 'correspondence to observable reality', instead of 'objectivity'

Science is practised by human beings, who have access to the real world only through interpretation of their perceptions. Taking objectivity seriously as an ideal, scientists need to make the sharing of their perceptions and interpretations possible. When applied to statistics, the implication is that choices in data analysis (including the prior distribution, if any, but also the model for the data, methodology and the choice of what information to include in the first place) should be motivated on the basis of factual, externally verifiable information and transparent criteria. This is similar to the idea of the concept of 'institutional decision analysis' (section 9.5 of Gelman *et al*. (2013)), under which the mathematics of formal decision theory can be used to ensure that decisions can be justified on the basis of clearly stated criteria. Different stakeholders will disagree on decision criteria, and different scientists will differ on statistical modelling decisions, so, in general, there is no unique 'objective' analysis, but we can aim at

communicating and justifying analyses in ways that support scrutiny and eventually consensus. Similar thoughts have motivated the slogan 'transparency is the new objectivity' in journalism (Weinberger, 2009).

In the context of statistical analysis, a key aspect of objectivity is therefore a process of *transparency*, in which the choices that are involved are justified on the basis of external, potentially verifiable sources or at least transparent considerations (ideally accompanied by sensitivity analyses if such considerations leave alternative options open), a sort of 'paper trail' leading from external information, through modelling assumptions and decisions about statistical analysis, all the way to inferences and decision recommendations. The current push of some journals to share data and computer code and the advent of tools to organize code and projects such as Github and version control better go in this direction. Transparency also comprises spelling out explicit and implicit assumptions about the data production, some of which may be unverifiable.

But transparency is not enough. Science aims at stable *consensus* in potentially free exchange (see Section 2.3), which is one reason that the current crisis of non-replication is taken so seriously in psychology (Yong, 2012). Transparency contributes to this building of consensus by allowing scholars to trace the sources and information that are used in statistical reasoning (Gelman and Basbøll, 2013). Furthermore, scientific consensus, as far as it deserves to be called 'objective', requires rationales, clear arguments and motivation, along with elucidation of how this relates to already existing knowledge. Following generally accepted rules and procedures counters the dependence of results on the personalities of individual researchers, although there is always a danger that such generally accepted rules and procedures are inappropriate or suboptimal for the specific situation at hand. For such reasons, one might question the inclusion of consensus as a virtue. Its importance stems from the impossibility to access observer-independent reality which means that exchange between observers is necessary to find out about what can be taken as real and stable. Consensus cannot be enforced; as a virtue it refers to behaviour that facilitates consensus.

In any case, consensus can only be achieved if researchers attempt to be *impartial* by taking into account competing perspectives, avoiding favouring prechosen hypotheses, and being open to criticism. In the context of epidemiology, Greenland (2012) proposed transparency and neutrality as replacements for objectivity.

Going on, the world outside the observer's mind plays a key role in usual concepts of objectivity, and as explained in Section 2.3 we see it as a major target of science. We acknowledge that the 'real world' is accessible to human beings through observation only, and that scientific observation and measurement cannot be independent of human preconceptions and theories. As statisticians we are concerned with making general statements based on systematized observations, and this makes *correspondence to observed reality* a core concern regarding objectivity. This is not meant to imply that empirical statements about observations are the only meaningful statements that can be made about reality; we think that scientific theories that cannot be verified (but can potentially be falsified) by observations are meaningful thought constructs, particularly because observations are never 'pure' and truly independent of thought constructs. Certainly in some cases the measurements, i.e. the observations that the statistician deals with, require critical scrutiny before discussing any statistical analysis of them; see Section 2.2.

Formal statistical methods contribute to objectivity as far as they contribute to the fulfilment of these *desiderata*, particularly by making procedures and their implied rationales transparent and unambiguous.

For example, Bayesian statistics is commonly characterized as 'subjective' by Bayesians and non-Bayesians alike. But, depending on how exactly prior distributions are interpreted and used (see Sections 5.3–5.5), they fulfil or aid some or all of the virtues that were listed above. Priors

make the researchers' prior point of view transparent; different approaches of interpreting them provide different rationales for consensus; 'objective Bayesians' (see Section 5.4) try to make them impartial; and if suitably interpreted (see Section 5.5) they can be properly grounded in observations.

3.2. 'Multiple perspectives' and 'context dependence', instead of 'subjectivity'

Science is normally seen as striving for objectivity, and therefore acknowledging subjectivity can be awkward. But, as noted above already, reality and the facts are accessible only through individual personal experiences. Different people bring different information and different viewpoints, and they will use scientific results in different ways. To enable clear communication and consensus, differing perspectives need to be acknowledged, which contributes to transparency and thus to objectivity. Therefore, subjectivity is important to the scientific process. Subjectivity is valuable in statistics in that it represents a way to incorporate the information coming from differing perspectives, which are the building blocks of scientific consensus.

We propose *awareness of multiple perspectives* and *context dependence* as key virtues making explicit the value of subjectivity. To the extent that subjectivity in statistics is a good thing, it is because information truly is dispersed, and, for any particular problem, different stakeholders have different goals. A counterproductive implication of the idea that science should be 'objective' is that there is a tendency in the communication of statistical analyses either to avoid or hide decisions that cannot be made in an automatic, seemingly 'objective' fashion by the available data. Given that all observations of reality depend on the perspective of an observer, interpreting science as striving for a unique ('objective') perspective is illusory. Multiple perspectives are a reality to be reckoned with and should not be hidden. Furthermore, by avoiding personal decisions, researchers often waste opportunities to adapt their analyses appropriately to the context, the specific background and their specific research aims, and to communicate their perspective more clearly. Therefore we see the acknowledgement of multiple perspectives and context dependence as virtues, making clearer in which sense subjectivity can be productive and helpful.

The term 'subjective' is often used to characterize aspects of certain statistical procedures that cannot be derived automatically from the data to be analysed, such as Bayesian prior distributions, tuning parameters (e.g. the proportion of trimmed observations in trimmed means, or the threshold in wavelet smoothing), or interpretations of data visualization. Such decisions are entry points for multiple perspectives and context dependence. The first decisions of this kind are typically the choice of data to be analysed and the family of statistical models to be fitted.

To connect with the other part of our proposal, the recognition of different perspectives should be done in a transparent way. We should not say that we set a tuning parameter to 2.5 (say) just because that is our belief. Rather, we should justify the choice explaining clearly how it supports the research aims. This could be by embedding the choice in a statistical model that can ultimately be linked back to observable reality and empirical data, or by reference to desirable characteristics (or avoidance of undesirable artefacts) of the methodology given the use of the chosen parameter; actually, many tuning parameters are related to such characteristics and aims of the analysis rather than to some assumed underlying 'belief' (see Section 4.3). In many cases, such a justification may be imprecise, for example because background knowledge may be only qualitative and not quantitative or not sufficiently precise to tell possible alternative choices apart, but often it can be argued that even then conscious tuning or specification of a prior distribution comes with benefits compared with using default methods of which the main attraction often is that seemingly 'subjective' decisions can be avoided.

To consider an important example, regularization requires such decisions. Default priors on regression coefficients are used to express the belief that coefficients are typically close to 0, and, from a non-Bayesian perspective, lasso shrinkage can be interpreted as encoding an external assumption of sparsity. Sparsity assumptions can be connected to an implicit or explicit model in which problems are in some sense being sampled from some distribution or probability measure of possible situations; see Section 5.5. This general perspective (which can be seen as Bayesian with an implicit prior on states of nature, or classical with an implicit reference set for the evaluation of statistical procedures) provides a potential basis to connect choices to experience; at least it makes transparent what kind of view of reality is encoded in the choices.

Tibshirani (2014) wrote that enforcing sparsity is not primarily motivated by beliefs about the world, but rather by benefits such as computability and interpretability, hinting at the fact that considerations other than being 'close to the real world' often play an important role in statistics and more generally in science. Even in areas such as social science where no underlying truly sparse structure exists, imposing sparsity can have advantages such as supporting stability (Gelman, 2013).

In a wider sense, if one is performing a linear or logistic regression, for example, and considering options of maximum likelihood, the lasso or hierarchical Bayes with a particular structure of priors, all of these choices are 'subjective' in the sense of encoding aims regarding possible outputs and assumptions, and all are 'objective' as far as these aims and assumptions are made transparent and the assumptions can be justified on the basis of past data and ultimately be checked given enough future data. So the conventional labelling of Bayesian analyses or regularized estimates as 'subjective' misses the point.

For another example, the binomial data confidence interval based on $(y+2)/(n+4)$ gives better coverage than the classical interval based on y/n (Agresti and Coull, 1998). Whereas the latter has a straightforward justification, the former is based on trading interval width against conservatism and involves some approximation and simplification, which the authors justified by the fact that the resulting formula can be presented in elementary courses. Debating whether this is more subjective than the classical approach, and whether this is a problem, is not helpful. Similarly, when comparing Bayesian estimates of public opinion by using multilevel regression and poststratification to taking raw survey means (which indeed correspond to Bayesian analyses under unreasonable flat priors), it is irrelevant which is considered more subjective.

Tuning parameters can be set or estimated on the basis of past data, and also on the basis of understanding of the effect of the choice on results and a clear explanation why a certain impact is desired or not. In robust statistics, for example, the breakdown point of some methods can be tuned and may be chosen lower than the optimal 50% because, if there is too large a percentage of data deviating strongly from the majority, one may rather want the method to deliver a compromise between all observations but, if the percentage of outliers is quite low, one may rather want them to be disregarded, with borderline percentages depending on the application (particularly on to what extent outliers are interpreted as erroneous observations rather than as somewhat special but still relevant cases).

Here is an example in which awareness of multiple perspectives can help with a problem with impartiality. Simulation studies for comparing statistical methods are often run by the designers of one of the competing approaches and, even if this is not so, the person running the study may have prior opinions about the competitors that may affect the study. There is simply no 'objective' way how this can be avoided; taking into account multiple perspectives by for example asking designers of all competing methods to provide simulation set-ups might help here.

3.3. Stability

As outlined in Section 2.3, we believe that science aims at a stable and reliable view of reality. Human beings do not have direct access to observer-independent reality, but phenomena that remain stable when perceived through different channels, at different times, and that are confirmed as stable by different observers, are the best contenders to be attributed 'objective existence'.

The term *stability* is difficult to find in philosophical accounts of objectivity, but it seems that Mayo's (1996) view of the growth of experimental knowledge through piecemeal testing of aspects of scientific theories and learning from error (which to her is a key feature of objectivity) implicitly aims at probing the stability of these theories. Stability is also connected to subjectivity in the sense that in the best case stability persists under inquiry from as many perspectives and in as many contexts as possible.

The accommodation and analysis of variability is something that statistical modelling brought to science, and in this sense statisticians investigate stability (of observations as well as of the statistics and estimators computed from them) all the time. An investigation of stability just based on variability assuming a parametric model is quite narrow, though, and there are many further sources of potential instabilities. Stability can refer to reproducibility of conclusions on new data, or to alternative analyses of the same data making different choices regarding for example tuning constants, Bayesian priors, transformations, resampling, removing outliers or even completely different methodology as far as this aims at investigating the same issue (alternative analyses that can be interpreted as doing something essentially different cannot be expected to deliver a similar result). On the most basic (but not always trivial) level, the same analysis on the same data should be replicable by different researchers. In statistical theory, basic variability assuming a parametric model can be augmented by robustness against various violations of model assumptions and Bayesian sensitivity analysis.

There are many aspects of stability that can be investigated, and only so much can be expected from a single study or publication; the generation of reliable scientific knowledge generally requires investigation of phenomena from more points of view than that of a single researcher or team.

3.4. A list of specific virtues

To summarize the above discussion, we give a more detailed list of the virtues that were discussed above, which we think will improve on discussions in which approaches, analyses and arguments are branded 'subjective' or 'objective' (Table 1).

In the subsequent discussion we refer to the item numbers in the list in Table 1 starting by V for virtue, such as V4(b) for 'clear conditions for reproduction, testing and falsification'.

We are aware that in some situations some of these virtues may oppose each other; for example 'consensus' can contradict 'awareness of multiple perspectives', and indeed dissent is essential to scientific progress. This tension between impersonal consensus and creative debate is an unavoidable aspect of science. Sometimes the consensus can only be that there are different legitimate points of view. Furthermore, the virtues listed are not all fully autonomous; clear reference to observations may be both a main rationale for consensus and a key contribution to transparency; and the subjective virtues contribute to both transparency and openness to criticism and exchange.

Not all items on the list apply to all situations. For example, in Section 5 we apply the list to the foundations of statistics, but some virtues (such as full communication of procedures) rather apply to specific studies.

Table 1. Virtues

V1. Transparency
(a) Clear and unambiguous definitions of concepts
(b) Open planning and following agreed protocols
(c) Full communication of reasoning, procedures, spelling out of (potentially unverifiable) assumptions and potential limitations
V2. Consensus
(a) Accounting for relevant knowledge and existing related work
(b) Following generally accepted rules where possible and reasonable
(c) Provision of rationales for consensus and unification
V3. Impartiality
(a) Thorough consideration of relevant and potentially competing theories and points of view
(b) Thorough consideration and if possible removal of potential biases: factors that may jeopardize consensus and the intended interpretation of results
(c) Openness to criticism and exchange
V4. Correspondence to observable reality
(a) Clear connection of concepts and models to observables
(b) Clear conditions for reproduction, testing and falsification
V5. Awareness of multiple perspectives
V6. Awareness of context dependence
(a) Recognition of dependence on specific contexts and aims
(b) Honest acknowledgement of the researcher's position, goals, experiences and subjective point of view
V7. Investigation of stability
(a) Consequences of alternative decisions and assumptions that could have been made in the analysis
(b) Variability and reproducibility of conclusions on new data

4. Applied examples

In conventional statistics, assumptions are commonly minimized. Classical statistics and econometrics are often framed in terms of robustness, with the goal being methods that work with minimal assumptions. But the decisions about what information to include and how to frame the model—these are typically buried, not stated formally as assumptions but just baldly stated: 'Here is the analysis we did . . .,' sometimes with the statement or implication that these have a theoretical basis but typically with little clear connection between subject matter theory and details of measurements. From the other perspective, Bayesian analyses are often boldly assumption based but with the implication that these assumptions, being subjective, need no justification and cannot be checked from data.

We would like statistical practice, Bayesian and otherwise, to move towards more transparency regarding the steps linking theory and data to models, and recognition of multiple perspectives in the information that is included in this paper trail and this model. In this section we show how we are trying to move in this direction in some of our recent research projects. We present these examples not as any sort of ideals but rather to demonstrate how we are grappling with these ideas and, in particular, the ways in which active awareness of the concepts of transparency, consensus, impartiality, correspondence to observable reality, multiple perspectives and context dependence is changing our applied work.

4.1. A hierarchical Bayesian model in pharmacology

Statistical inference in pharmacokinetics–pharmacodynamics involves many challenges: data are indirect and often noisy; the mathematical models are non-linear and computationally expensive, requiring the solution of differential equations; parameters vary by person but often with only a small amount of data on each experimental subject. Hierarchical models and Bayesian inference are often used to manage the many levels of variation and uncertainty (see, for example, Sheiner (1984) and Gelman *et al*. (1996)).

One of us is currently working on a project in drug development involving a Bayesian model that was difficult to fit, even when using advanced statistical algorithms and software. Following the so-called folk theorem of statistical computing (Gelman, 2008), we suspected that the problems with computing could be attributed to a problem with our statistical model. In this case, the issue did not seem to be a lack of fit, or a missing interaction, or unmodelled measurement error—problems we had seen in other settings of this sort. Rather, the fit appeared to be insufficiently constrained, with the Bayesian fitting algorithm being stuck going through remote regions of parameter space that corresponded to implausible or unphysical parameter values.

In short, the model as written was only weakly identified, and the given data and priors were consistent with all sorts of parameter values that did not make scientific sense. Our iterative Bayesian computation had poor convergence—i.e. the algorithm was having difficulty approximating the posterior distribution—and the simulations were going through zones of parameter space that were not consistent with the scientific understanding of our pharmacology colleagues.

To put it another way, our research team had access to prior information that had not been included in the model. So we took the time to specify a more informative prior. The initial model thus played the role of a placeholder or default which could be elaborated as needed, following the iterative prescription of falsificationist Bayesianism (Box (1980) and Gelman *et al*. (2013), section 5.5).

In our experience, informative priors are not so common in applied Bayesian inference and, when they are used, they often seem to be presented without clear justification. In this instance, though, we decided to follow the principle of transparency and to write a note explaining the genesis of each prior distribution. To give a sense of what we are talking about, we present a subset of these notes here:

'• γ_1: mean of population distribution of $\log(\text{BVA}_j^{\text{latent}}/50)$, centered at 0 because the mean of the BVA values in the population should indeed be near 50. We set the prior sd to 0.2 which is close to $\log(60/50) = 0.18$ to indicate that we're pretty sure the mean is between 40 and 60.

- γ_2: mean of pop dist of $\log(k_j^{\text{in}}/k_j^{\text{out}})$, centered at 3.7 because we started with -2.1 for k^{in} and -5.9 for k^{out}, specified from the literature about the disease. We use a sd of 0.5 to represent a certain amount of ignorance: we're saying that our prior guess for the population mean of $k^{\text{in}}/k^{\text{out}}$ could easily be off by a factor of $\exp(0.5) = 1.6$.
- γ_3: mean of pop dist of $\log k_j^{\text{out}}$, centered at -5.8 with a sd of 0.8, which is the prior that we were given before, from the time scale of the natural disease progression.
- γ_4: $\log E_{\text{max}}^0$, centered at 0 with sd 2.0 because that's what we were given earlier.'

The γs here already represent a transformation of the original parameters, BVA (baseline visual acuity; this is a drug for treating vision problems), k^{in} and k^{out} (rate constants for differential equations that model the diffusion of the drug) and E_{max}^0, a saturation parameter in the model. One goal in this sort of work is to reparameterize to unbounded scales (so that normal distributions are more reasonable, and we can specify parameters based on location and scale) and to aim for approximate independence in the prior distribution because of the practical difficulties of eliciting prior correlations. The 'literature about the disease' comes from previously published trials of other drugs for this disease; these trials also include control arms which give us information on the natural progression of visual acuity in the absence of any treatment.

We see this sort of painfully honest justification as a template for future Bayesian data analyses. The above snippet certainly does not represent an exemplar of best practices, but we see it as a 'good enough' effort that presents our modelling decisions in the context in which they were made.

To label this prior specification as 'objective' or 'subjective' would miss the point. Rather, we see it as having some of the virtues of objectivity and subjectivity—notably, transparency (virtue V1) and some aspects of consensus (virtue V2) and awareness of multiple perspectives (virtue V5)—while recognizing its clear imperfections and incompleteness. Other desirable features would derive from other aspects of the statistical analysis—for example, we use external validation to approach correspondence to observable reality (virtue V4), and our awareness of context dependence (virtue V6) comes from the placement of our analysis within the larger goal, which is to model dosing options for a particular drug.

One concern about our analysis which we have not yet thoroughly addressed is sensitivity to model assumptions. We have established that the prior distribution makes a difference but it is possible that different reasonable priors yield posteriors with greatly differing real world implications, which would raise concern about consensus (virtue V2) and impartiality (virtue V3). Our response to such concerns, if this sensitivity is indeed a problem, would be to document our choice of prior more carefully, thus doubling down on the principle of transparency (virtue V1) and to compare with other possible prior distributions supported by other information, thus supporting impartiality (virtue V3) and awareness of multiple perspectives (virtue V5).

The point is not that our particular choices of prior distributions are 'correct' (whatever that means) or optimal, or even good, but rather that they are transparent, and in a transparent way connected to knowledge. Subsequent researchers—whether supportive, critical or neutral regarding our methods and substantive findings—should be able to interpret our priors (and, by implication, our posterior inferences) as the result of some systematic process, a process which is sufficiently open that it can be criticized and improved as appropriate.

4.2. Adjustments for pre-election polls

Wang *et al*. (2014) described another of our recent applied Bayesian research projects, in this case a statistical analysis that allows highly stable estimates of public opinion by adjustment of data from non-random samples. The particular example that was used was an analysis of data from an opt-in survey conducted on the Microsoft Xbox video game platform, a technique that allowed the research team, effectively to interview respondents in their living rooms, without ever needing to call or enter their houses.

The Xbox survey was performed during the two months before the 2012 US presidential election. In addition to offering the potential practical benefits of performing a national survey using inexpensive data, this particular project made use of its large sample size and panel structure (repeated responses on many thousands of Americans) to learn something new about US politics: we found that certain swings in the polls, which had been generally interpreted as representing large swings in public opinion, actually could be attributed to differential non-response, with Democrats and Republicans in turn being more or less likely to respond during periods where there was good or bad news about their candidate. This finding was consistent with some of the literature in political science (see Erikson *et al*. (2004)), but the Xbox study represented an important empirical confirmation (Gelman *et al*., 2016).

Having established the potential importance of the work, we next consider its controversial aspects. For many decades, the gold standard in public opinion research has been probability

sampling, in which the people being surveyed are selected at random from a list or lists (e.g. selecting households at random from a list of addresses or telephone numbers and then selecting a person within each sampled household from a list of the adults who live there). From this standpoint, opt-in sampling of the sort that was employed in the Xbox survey lacks a theoretical foundation, and the estimates and standard errors thus obtained (and which we reported in our research papers) do not have a clear statistical interpretation.

This criticism—that inferences from opt-in surveys lack a theoretical foundation—is interesting to us here because it is *not* framed in terms of objectivity or subjectivity. We do use Bayesian methods for our survey adjustment but the criticism from certain survey practitioners is not about adjustment but rather about the data collection: they take the position that no good adjustment is possible for data that are collected from a non-probability sample.

As a practical matter, our response to this criticism is that non-response rates in national random-digit-dialled telephone polls are currently in the range of 90%, which implies that real world surveys of this sort are essentially opt-in samples in any case: if there is no theoretical justification for non-random samples then we are left with the choice either to abandon statistical inference entirely when dealing with survey data, or to accept that our inferences are model based and to do our best (Gelman, 2014c).

Our Bayesian adjustment model (Wang *et al*., 2014) used prior information in two ways. First, population distributions of demographics, state of residence and party identification were imputed by using exit poll data from the previous election; from the survey sampling perspective this was a poststratification step, and from the political science perspective this represents an assumption of stability in the electorate from 2008 to 2012. The second aspect of prior information was encoded in our hierarchical logistic regression model, in which varying intercepts for states and for different demographic factors were modelled as exchangeable batches of parameters drawn from normal distributions. These assumptions are necessarily approximate and are thus ultimately justified on pragmatic grounds.

We shall now express this discussion by using the criteria from Section 4. Probability sampling has the clear advantage of transparency (virtue V1) in that the population and sampling mechanism can be clearly defined and accessible to outsiders, in a way that an opt-in survey such as the Xbox survey is not. In addition, the probability sampling has the benefits of consensus (virtue V2), at least in the USA, where such surveys have a long history and are accepted in marketing and opinion research. Impartiality (virtue V3) and correspondence to observable reality (virtue V4) are less clearly present because of the concern with non-response, just noted. We would argue that the large sample size and repeated measurements of the Xbox data, coupled with our sophisticated hierarchical Bayesian adjustment scheme, put us well on the road to impartiality (through the use of multiple sources of information, including past election outcomes, used to correct for biases in the form of known differences between sample and observation) and correspondence to observable reality (in that the method can be used to estimate population quantities that could be validated from other sources).

Regarding the virtues that are associated with subjectivity, the various adjustment schemes represent awareness of context dependence (virtue V6) in that the choice of variables to match in the population depend on the context of political polling, both in the sense of which aspects of the population are particularly relevant for this purpose, and in respecting the awareness of survey practitioners of what variables are predictive of non-response. The researcher's subjective point of view is involved in the choice of exactly what information to include in weighting adjustments and exactly what statistical model to fit in regression-based adjustment. Users of probability sampling on grounds of 'objectivity' may shrink from using such judgements and may therefore ignore valuable information from the context.

4.3. Transformation of variables in cluster analysis for socio-economic stratification

Cluster analysis aims at grouping similar objects and separating dissimilar objects, and as such is based, explicitly or implicitly, on some measure of dissimilarity. Defining such a measure, e.g. by using some set of variables characterizing the objects to be clustered, can involve many decisions. Here we consider an example of Hennig and Liao (2013), where we clustered data from the 2007 US Consumer Finances Survey, comprising variables on income, savings, housing, education, occupation, number of checking and savings accounts, and life insurance with the aim of data-based exploration of socio-economic stratification. The choice of variables and the decisions of how they are selected, transformed, standardized and weighted has a strong effect on the results of the cluster analysis. This effect depends to some extent on the clustering technique that is afterwards applied to the resulting dissimilarities but will typically be considerable, even for cluster analysis techniques that are not directly based on dissimilarities. One of the various issues that were discussed by Hennig and Liao (2013) was the transformation of the variables treated as continuous (namely income and savings amount), with the view of basing a cluster analysis on a Euclidean distance after transformation, standardization and weighting of variables.

There is some literature on choosing transformations, but the usual aims of transformation, namely achieving approximate additivity, linearity, equal variances or normality, are often not relevant for cluster analysis, where such assumptions apply only to model-based clustering, and only within the clusters, which are not known before transformation.

The rationale for transformation when setting up a dissimilarity measure for clustering is of a different kind. The measure needs to formalize appropriately which objects are to be treated as 'similar' or 'dissimilar' by the clustering methods and should therefore be put into the same or different clusters respectively. In other words, the formal dissimilarity between objects should match what could be called the 'interpretative dissimilarity' between objects. This is an issue involving subject matter knowledge that cannot be decided by the data alone.

Hennig and Liao (2013) argued that the interpretative dissimilarity between different savings amounts is governed rather by ratios than by differences, so that \$2 million of savings is seen as about as dissimilar from \$1 million, as \$2000 is dissimilar from \$1000. This implies a logarithmic transformation. We do not argue that there is a precise argument that privileges the log-transformation over other transformations that achieve something similar, and one might argue from intuition that even taking logarithms may not be sufficiently strong. We therefore recognize that any choice of transformation is a provisional device and only an approximation to an ideal 'interpretative dissimilarity', even if such an ideal exists.

In the data set, there are no negative savings values as there is no information on debts, but there are many people who report zero savings, and it is conventional to kludge the logarithmic transformation to become $x \mapsto \log(x+c)$ with some $c > 0$. Hennig and Liao (2013) then pointed out that, in this example, the choice of c has a considerable effect on clustering. The number of people with very low but non-zero savings in the data set is quite small. Setting $c = 1$, for example, the transformation creates a substantial gap between the zero-savings group and people with fairly low (but non-zero) amounts of savings, and of course this choice is also sensitive to scaling (for example, savings might be coded in dollars, or in thousands of dollars). The subsequent cluster analysis (done by 'partitioning around medoids'; Kaufman and Rousseeuw (1990)) would therefore separate the zero-savings group strictly; no person with zero savings would appear together in a cluster with a person with non-zero savings. For larger values for c, the dissimilarity between the zero-savings group and people with a low savings amount becomes effectively sufficiently small that people with zero savings could appear in clusters together with other people, as long as values on other variables are sufficiently similar.

We do not believe that there is a true value of c. Rather, clusterings arising from different choices of c are legitimate but imply different interpretations. The clustering for $c = 1$ is based on treating the zero-savings group as special, whereas the clustering for $c = 200$, say, implies that a difference in savings between \$0 and \$100 is taken as not so important (although it is more important in any case than the difference between \$100 and \$200). Similar considerations hold for issues such as selecting and weighting variables and coding ordinal variables.

It can be frustrating to the novice in cluster analysis that such decisions for which there do not seem to be an objective basis can make such a difference, and there is apparently a strong temptation to ignore the issue and just to choose $c = 1$, which may look natural in the sense that it maps zero onto zero, or even to avoid transformation at all to avoid the discussion, so that no obvious lack of objectivity strikes the reader. Having the aim of socio-economic stratification in mind, though, it is easy to argue that clusterings that result from ignoring the issue are less desirable and useful than a clustering obtained from making a however imprecisely grounded decision choosing $c > 1$, therefore avoiding either separation of the zero-savings group as a clustering artefact or an undue domination of the clustering by people with large savings in case of not applying any transformation at all.

We believe that this kind of tuning problem that cannot be interpreted as estimating an unknown true constant (and does therefore not lend itself naturally to an approach through a Bayesian prior) is not exclusive to cluster analysis and is often hidden in presentations of data analyses.

Hennig and Liao (2013) pointed out the issue and did some sensitivity analysis about the strength of the effect of the choice of c (virtue V7a). The way that we picked c in Hennig and Liao (2013) made clear reference to the context dependence, while being honest that the subject matter knowledge in this case provided only weak guidelines for making this decision (virtue V6). We were also clear that alternative choices would amount to alternative perspectives rather than being just wrong (virtues V5 and V3).

The issue how to foster consensus and to make a connection to observable reality (virtues V2 and V4) is of interest but is not treated here.

But it is problematic to establish rationales for consensus that are based on ignoring context and potentially multiple perspectives. There is a tendency in the cluster analysis literature to seek formal arguments for making such decisions automatically (see, for example, Everitt *et al*. (2011), section 3.7, on variable weighting; it is difficult to find anything systematic in the clustering literature on transformations), trying to optimize 'clusterability' of the data set, or preferring methods that are less sensitive to such decisions, because this amounts to making the decisions implicitly without giving the researchers access to them. In other words, the data are given the authority to determine not only which objects are similar (which is what we want them to do), but also what similarity should mean. The latter should be left to the researcher, although we acknowledge that the data can have a certain influence: for example the idea that dissimilarity of savings amounts is governed by ratios rather than differences is connected to (but not determined by) the fact that the distribution of savings amounts is skewed, with large savings amounts sparsely distributed.

4.4. Testing for homogeneity against clustering

An issue in Hennig and Liao (2013) was whether there is any meaningful clustering to be found in the data. Some sociologists suspect that, in many modern democratic societies, stratification may represent no more than a drawing of arbitrary borders through a continuum of socio-economic conditions. We were interested in what the data have to say on this issue, and we

chose to address this by running a test of a homogeneity null hypothesis against a clustering alternative (knowing that there is some distance to go between the result of such an analysis and the 'desired' sociological interpretation).

If we had been concerned primarily with appearing objective and the ease to achieve a significant result, probably we would have chosen a likelihood ratio test of the null hypothesis of a standard homogeneity model (in the specific situation this could have been a Gaussian distribution for the continuous variables, an uncorrelated adjacent category ordinal logit model for ordinal variables and a locally independent multinomial model for categorical data) for a single mixture component in the framework of mixture models as provided, for example, in the LatentGOLD software package (Vermunt and Magidson, 2016).

But, even in the absence of meaningful clusters, real data do not follow such clean distributional shapes and therefore sufficiently large data sets (including ours, with $n > 17000$) will almost always reject a simple homogeneity model. We therefore set out to build a null model that captured the features of the data set such as the dependence between variables and marginal distributions of the categorical variables as well as possible, without involving anything that could be interpreted as clustering structure. As opposed to the categorical variables, the marginal distributions of the 'continuous' variables such as the transformed savings amount were treated as potentially indicating clustering, and therefore the null model used non-parametric unimodal distributions for them. Data from this null model involving several characteristics estimated from the data could be simulated by using the parametric bootstrap.

As test statistic we used a cluster validity statistic of the clustering computed on the data, which was not model based but dissimilarity based. The idea behind this was that we wanted a test statistic which would measure the degree of clustering, so that we could find out how much 'clustering' we could expect to see even if no meaningful clustering was present (under the null model). Actually we computed clusterings for various numbers of clusters. Rather than somehow to define a single p-value from aggregating all these clusterings (or selecting the 'best' one), we decided to show a plot of the values of the validity statistic for the different numbers of clusters for the real data set together with the corresponding results for many data sets simulated from the null model. The result of this showed clearly that a higher level of clustering was found in the real data set.

In doing this, we deviated from classical significance test logic in several ways, by not using a test statistic that was optimal against any specific alternative, by not arguing from a single p-value and by using a null model that relied heavily on the data to try as hard as we can to model the data without clustering. Still, in case that the validity statistic values for the real data do not look clearly different from those of the bootstrapped data set, this can be interpreted as no evidence in the data for real clustering, whereas the interpretation of a clear ('significant') difference depends on whether we can argue convincingly that the null model is as good as possible at trying to model the data without clustering structure. Setting up a straw man null model for homogeneity and rejecting it would have been easy and not informative. The general principle is discussed in more detail in Hennig and Lin (2015), including real data examples where such a null model could not be rejected, as opposed to a straw man model.

The essence here is that we made quite a number of decisions that opened our analysis more clearly to the charge of 'not being objective' than following a standard approach, for the sake of adapting the analysis better to the specific data in hand, and of giving the null hypothesis the best possible chance (the non-rejection of it would have been a non-discovery here; the role of it was not to be 'accepted' as 'true' anyway).

We tried to do good science, though, by checking as impartially and transparently as we could (virtues V1 and V3), whether the data support the idea of a real clustering (virtue V4).

This involved context-dependent judgement (virtue V6) and the transparent choice of a specific perspective (the chosen validity index) among a potential variety (virtue V5), because we were after more qualitative statements than degrees of belief in certain models.

5. Decomposing subjectivity and objectivity in the foundations of statistics

In this section, we use the above list of virtues to revisit aspects of the discussion on fundamental approaches to statistics, for which the terms 'subjective' and 'objective' typically play a dominant role. We discuss what we perceive to be the major streams of the foundations of statistics, but within each of these streams there are several approaches, which we cannot cover completely in such a paper; rather we sketch the streams somewhat roughly and refer to only a single or a few leading references for details where needed.

Here, we distinguish between interpretations of probability, and approaches for statistical inference. For example, 'frequentism' as an interpretation of probability does not necessarily imply that Fisherian or Neyman–Pearson tests are preferred to Bayesian methods, despite the fact that frequentism is more often associated with the former than with the latter.

We shall go through several philosophies of statistical inference, for each laying out the connections that we see to the virtues of objectivity and subjectivity outlined in Section 3.4.

Exercising awareness of multiple perspectives, we emphasize that we do not believe that one of these philosophies is the correct or best; nor do we claim that reducing the different approaches to a single approach would be desirable. What is lacking here is not unification, but rather, often, transparency about which interpretation of probabilistic outcomes is intended when applying statistical modelling to specific problems. Particularly, we think that, depending on the situation, both 'aleatory' or 'epistemic' approaches to modelling uncertainty are legitimate and worthwhile, referring to data-generating processes in observer-independent reality on one hand and rational degrees of belief on the other.

We focus on approaches that are conventionally labelled as either Bayesian or frequentist, but we acknowledge that there are important perspectives on statistics that lie outside this traditional divide. Discussing them in detail would be worthwhile but is beyond our focus, and we hope that discussants of our paper will pick up these threads. Examples of other perspectives include *machine learning*, where the focus is on prediction rather than parameter estimation; thus there is more emphasis on correspondence to observable reality (virtue V4) compared with other virtues, *alternative models of uncertainty* such as belief functions, imprecise probabilities, and fuzzy logic that aim to circumvent some of the limitations of probability theory (most notoriously, the difficulty of distinguishing between 'known unknowns' and 'unknown unknowns', or risk and uncertainty in the terminology of Knight (1921)) and *exploratory data analysis* (Tukey, 1977), which is sensitive to multiple perspectives (virtue V5) and context dependence (virtue V6), and tries to be more directly connected to the data than if it was mediated by probability models (virtue V4(a)). Whether avoidance of probability modelling contributes to transparency (virtue V1(a)) is rather problematic because implicit assumptions that may not be spelled out (virtue V1(c)) can be controversial.

5.1. Frequentist probabilities

'Frequentism' as an interpretation of probability refers, in a narrow sense, to the identification of the probability of an event in a certain experiment with a limiting relative frequency of occurrences if the experiment were to be carried out infinitely often in some kind of independent manner. Frequentist statistics is based on evaluating procedures based on a long-term average over a 'reference set' of hypothetical replicated data sets. Different choices of reference sets were

for example used by Fisher (1955) and Pearson (1955) when discussing permutation or χ^2-tests for 2×2 tables.

In the wider sense, we call probabilities 'frequentist' when they formalize observer-independent tendencies or propensities of experiments to yield certain outcomes (see, for example, Gillies (2000)), which are thought of as replicable and yielding a behaviour under infinite replication as suggested by what is assumed to be the 'true' probability model.

The frequentist mindset locates probabilities in the observer-independent world, so they are in this sense objective (and often called 'objective' in the literature, e.g. Kendall (1949)). This, however, does not guarantee that frequentist probabilities really exist; an infinite number of replicates cannot exist, and even a finite amount of real replicates will neither be perfectly identical nor perfectly independent. Ultimately the ideally infinite populations of replicates are constructed by the 'statistician's imagination' (Fisher, 1955).

The decision to adopt the frequentist interpretation of probability regarding a certain phenomenon therefore requires idealization. It cannot be enforced by observation; nor is there generally enough consensus that this interpretation applies to any specific set-up, although it is well discussed and supported in some physical settings such as radioactive decay (virtues V2 and V4). Once a frequentist model has been adopted, however, it makes predictions about observations that can be checked, so the reference to the observable reality (virtue V4) is clear.

There is some disagreement about whether the frequentist definition of probability is clear and unambiguous (virtue V1(a)). On one hand, the idea of a tendency of an experiment to produce certain outcomes as manifested in observed and expected relative frequencies seems quite clear. On the other hand, it is difficult to avoid the circularity that would result from referring to independent and identical replicates when defining frequentist probabilities, because the standard definition of the terms 'independent' and 'identical' assumes a definition of probability that is already in place (see von Mises (1957) for a prominent attempt to solve this, and Fine (1973) for a criticism).

Frequentism implies that, in the observer-independent reality, true probabilities are unique, but there is considerable room for multiple perspectives (virtue V5) regarding the definition of replicable experiments, collectives or reference sets. The idea of replication is often constructed in a rather creative way. For example, in time series modelling the frequentist interpretation implies an underlying true distribution for every single time point, but there is no way to repeat observations independently at the same time point. This actually means that the effective sample size for time series data would be 1, if replication were not implicitly constructed in the statistical model, e.g. by assuming independent innovations in auto-regressive moving average type models. Such models, or, more precisely, certain aspects of such models, can be checked against the data but, even if such a check does not fail, it is still clear that there is no such thing in observable reality, even approximately, as a marginal 'true' frequentist distribution of the value of the time series x_t at fixed t, as implied by the model, because x_t is strictly not replicable.

The issue that useful statistical models require a construction of replication (or exchangeability) on some level by the statistician is, as we discuss below, not confined to frequentist models. To provide a rationale for the essential statistical task of pooling information from many observations to make inference relevant for future observations, all these observations need to be assumed to represent the same process somehow.

The appropriateness of such assumptions in a specific situation can often only be tested in quite a limited way by observations. All kinds of informal arguments can apply about why it is a good or bad idea to consider a certain set of observations (or unobservable implied entities such as error terms and latent variables) as independent and identically distributed frequentist replicates.

Unfortunately, although such an openness to multiple perspectives and potential context dependence (virtue V6(a)) can be seen as positive from our perspective, those issues involved in the choices of a frequentist reference set are often not clearly communicated and discussed. The existence of a true model with implied reference set is typically taken for granted by frequentists, motivated at least in part by the desire for objectivity.

5.2. Frequentist inference

This section is about inference from data about characteristics of an assumed true frequentist probability model. Traditionally, this comprises hypothesis tests, confidence intervals, and parameter estimators, but is not limited to them; see below.

According to Mayo and Spanos (2010) and Cox and Mayo (2010), a fundamental feature of frequentist inference is the evaluation of error probabilities, i.e. probabilities of wrong decisions. Traditionally these would be the type I and type II errors of Neyman–Pearson hypothesis testing, but the the error statistical perspective could also apply to other constructs such as errors of sign and magnitude ('type S' and 'type M' errors; Gelman and Carlin (2014)).

Mayo and co-workers see the ability to learn from error and to test models severely (in such a way that it would be difficult for a model to pass a test if it was wrong regarding the specific aspect that is assessed by a test) against data as a major feature of objectivity, which is made possible by the frequentist interpretation of probability measures as 'data generators'. In our list of virtues, this feature is captured in virtue V4(b) (reference to observations: reproduction; testing; falsification). The underlying idea, with which we agree, is that learning from error is a main driving force in science: a lifetime contract between the mode of statistical investigation and its object. This corresponds to Chang's active scientific realism that was mentioned above.

The error probability characteristics of methods for frequentist inference rely, in general, on model assumptions. In principle, these assumptions can be tested, also, and are therefore, according to Mayo and co-workers, no threat to the objectivity of the account. But this comes with two problems. Firstly, derivations of statistical inference based on error probabilities typically assume the model as fixed and do not account for prior model selection based on the data. This issue has recently attracted some research (e.g. Berk *et al*. (2013)), but this still requires a transparent listing of all the possible modelling decisions that could be made (virtue V1(b)), which often is missing, and which may not even be desirable as long as the methods are used in an exploratory fashion (Gelman and Loken, 2014). Secondly, any data set can be consistent with many models, which can lead to divergent inferences. Davies (2014) illustrates this with the analysis of a data set on amounts of copper in drinking water, which can be fitted well by a Gaussian, a double-exponential or a comb distribution, but yields vastly different confidence intervals for the centre of symmetry (which is assumed to be the target of inference) under these three models.

Davies (2014) suggested that it is misleading to hypothesize models or parameters to be 'true'. According to Davies, statistical modelling is about approximating the data in the sense that 'adequate' models are not rejected by tests based on characteristics of the data that the statistician is interested in (allowing for multiple perspectives and context dependence: virtues V5 and V6), i.e. they generate data that 'look like' the observed data with respect to the chosen characteristics. Regarding these characteristics, according to Davies, there is no essential difference between parameter values and distributional shapes or structural assumptions, and therefore no conceptual separation as in traditional frequentist inference between checking model assumptions and inference about parameters assuming a parametric model to be true. Such an approach is tied to the observations in a more direct way without making metaphysical assumptions about

unobservable features of observer-independent reality (virtues V1(a) and V4). It is frequentist inference in the sense that the probability models are interpreted as 'data generators'.

Two further streams in frequentist inference are concerned about the restrictiveness of parametric model assumptions. Robust statistics explores the stability (virtue V7) of inferences in case the 'true' model is not equal to the nominal model, but rather in some neighbourhood, and strives to develop methods that are stable in this respect. There are various ways to define such neighbourhoods and to measure robustness, so robustness considerations can bring in multiple perspectives (virtue V5) but may cause problems with reaching consensus (virtue V2).

Non-parametric statistics allows us to remove bias (virtue V3(c)) by minimizing assumptions regarding, for example, distributional shapes (structural assumptions such as independence are still required). In some cases, particularly with small data sets, this must be afforded by decreased stability (virtue V7).

Overall, there is no shortage of entry points for multiple perspectives (virtue V5) in frequentist inference. This could be seen as something positive, but it runs counter to some extent to the way that the approach is advertised as objective by some of its proponents. Many frequentist analyses could in our opinion benefit from acknowledging honestly their flexibility and the researcher's choices made, many of which cannot be determined by data alone.

5.3. Subjectivist Bayesianism

We call 'subjectivist epistemic' the interpretation of probabilities as quantifications of strengths of belief of an individual, where probabilities can be interpreted as derived from, or implementable through, bets that are coherent in that no opponent can cause sure losses by setting up some combinations of bets. From this requirement of coherence, the usual probability axioms follow (virtue V2(c)). Allowing conditional bets implies Bayes's theorem, and therefore, as far as inference concerns learning from observations about not (yet) observed hypotheses, Bayesian methodology is used for subjectivist epistemic probabilities: hence the term 'subjectivist Bayesianism'.

A major proponent of subjectivist Bayesianism was de Finetti (1974). de Finetti was not against objectivity in general. He viewed observed facts as objective, as well as mathematics and logic and certain formal conditions of random experiments such as the set of possible outcomes. But he viewed uncertainty as something subjective and he held that objective (frequentist) probabilities do not exist. He claimed that his subjectivist Bayesianism appropriately takes into account both the objective (see above) and subjective (opinions about unknown facts based on known evidence) components for probability evaluation.

In de Finetti's work the term 'prior' refers to all probability assignments using information that is external to the data at hand, with no fundamental distinction between the 'parameter prior' assigned to parameters in a model, and the form of the 'sampling distribution' given a fixed parameter, in contrast with common Bayesian practice today, in which the term 'prior' is used to refer only to the parameter prior. In the following discussion we shall use the term 'priors' in de Finetti's general sense.

Regarding the list of virtues in Table 1 in Section 3.4, de Finetti provided a clear definition of probability (virtue V1(a)) based on principles that he sought to establish as generally acceptable (virtue V2(c)). Unlike objectivist Bayesians, subjectivist Bayesians do not attempt to enforce agreement regarding prior distributions, not even given the same evidence; still, de Finetti (1974) and other subjectivist Bayesians proposed rational principles for assigning prior probabilities. There is also some work on (partial) intersubjective agreement on prior specifications, e.g. Dawid (1982a), providing a rationale for consensus (virtue V2(c)). The difference between the objectivist and subjectivist Bayesian point of view is rooted in the general tension in science that

was explained above; the subjectivist approach can be criticized for not supporting agreement sufficently—conclusions based on one prior may be seen as irrelevant for somebody who holds another (virtue V2(c))—but can be defended for honestly acknowledging that prior information often does not come in ways that allow a unique formalization (virtue V6(b)). In any case it is vital that subjectivist Bayesians explain transparently how they arrive at their priors, so that other researchers can decide to what extent they can support the conclusions (virtue V1(c)).

In de Finetti's conception, probability assessments, prior and posterior, can ultimately only concern observable events, because bets can only be evaluated if the experiment on which a bet is placed has an observable outcome, and so there is a clear connection to observables (virtue V4(a)).

However, priors in the subjectivist Bayesian conception are not open to falsification (virtue V4(b)), because by definition they must be fixed before observation. Adjusting the prior after having observed the data to be analysed violates coherence. The Bayesian system as derived from axioms such as coherence (as well as those used by objectivist Bayesians; see Section 5.4) is designed to cover all aspects of learning from data, including model selection and rejection, but this requires that all potential later decisions are already incorporated in the prior, which itself is not interpreted as a testable statement about yet unknown observations. In particular this means that, once a coherent subjectivist Bayesian has assessed a set-up as exchangeable *a priori*, he or she cannot drop this assumption later, whatever the data are (think of observing 20 0s, then 20 1s, and then 10 further 0s in a binary experiment). This is a major problem, because subjectivist Bayesians use de Finetti's theorem to justify working with parameter priors and sampling models under the assumption of exchangeability, which is commonplace in Bayesian statistics. Dawid (1982b) discussed calibration (quality of match between predictive probabilities and the frequency of predicted events to happen) of subjectivist Bayesians inferences, and he suggested that badly calibrated Bayesians could do well to adjust their future priors if this is needed to improve calibration, even at the cost of violating coherence.

Subjectivist Bayesianism scores well on virtues V5 and V6(b). But it is a limitation that the prior distribution exclusively formalizes belief; context and aims of the analysis do not enter unless they have implications about belief. In practice, an exhaustive elicitation of beliefs is rarely feasible, and mathematical and computational convenience often plays a role in setting up subjective priors, despite de Finetti's having famously accused frequentists of 'adhockeries for mathematical convenience'. Furthermore, the assumption of exchangeability will hardly ever precisely match an individual's beliefs in any situation—even if there is no specific reason against exchangeability in a specific set-up, the implicit commitment to stick to it whatever will be observed seems too strong—but some kind of exchangeability assumption is required by Bayesians for the same reason for which frequentists need to rely on independence assumptions: some internal replication in the model is needed to allow generalization or extrapolation to future observations; see Section 5.1.

Summarizing, we view much of de Finetti's criticism of frequentism as legitimate, and subjectivist Bayesianism comes with a commendable honesty about the effect of subjective decisions and allows for flexibility accommodating multiple perspectives. But checking and falsification of the prior are not built into the approach, and this can obstruct agreement between observers.

5.4. Objectivist Bayesianism

Given the way that objectivity is often advertised as a key scientific virtue (often without specifying what exactly it means), it is not surprising that de Finetti's emphasis on subjectivity is not shared by all Bayesians, and that there have been many attempts to specify prior distributions in a more objective way. Currently the approach of Jaynes (2003) seems to be among the most

popular. As with many of his predecessors such as Jeffreys and Carnap, Jaynes saw probability as a generalization of binary logic to uncertain propositions. Cox (1961) proved that, given a certain list of supposedly commonsense *desiderata* for a 'plausibility' measurement, all such measurements are equivalent, after suitable scaling, to probability measures. This theorem is the basis of Jaynes's objectivist Bayesianism, and the claim to objectivity comes from postulating that, given the same information, everybody should come to the same conclusions regarding plausibilities: prior and posterior probabilities (virtue V2(c)), a statement with which subjectivist Bayesians disagree.

In practice, this objectivist ideal seems to be difficult to achieve, and Jaynes (2003) admitted that setting up objective priors including all information is an unsolved problem. One may wonder whether his ideal is achievable at all. For example, in chapter 21, he gave a full Bayesian 'solution' to the problem of dealing with and identifying outliers, which assumes that prior models must be specified for both 'good' and 'bad' data (between which therefore there must be a proper distinction), including parameter priors for both models, as well as a prior probability for any number of observations to be 'bad'. It is difficult to see, and no information about this was provided by Jaynes himself, how it can be possible to translate the unspecific information of knowing of some outliers in many kinds of situations, some of which are more or less related, but none identical (say) to the problem at hand, into precise quantitative specifications as needed for Jaynes's approach in an objective way, all before seeing the data.

Setting aside the difficulties of working with informally specified prior information, a key issue of objectivist Bayesianism is the specification of an objective prior distribution formalizing the absence of information. Various principles for doing this have been proposed (maximum entropy, Jaynes (2003); maximum missing Shannon information, Berger *et al*. (2009); and a set of desirable properties, Bayarri *et al*. (2012)). Such principles have their difficulties and disagree in many cases (Kass and Wasserman, 1996). Objectivity seems to be an ambition rather than a description of what indeed can be achieved by setting up objectivist Bayesian priors. More modestly, therefore, Berger *et al*. (2009) used the term 'reference priors', avoiding the term 'objective', and emphasizing that it would be desirable to have a convention for such cases (virtue V2(b)), but admitting that it may not be possible to prove any general approach for arriving at such a convention uniquely correct or optimal in any rational sense. However, the proposal and discussion of such principles certainly served transparency (virtues V1(a) and V1(c)) and provided rationales for consensus (virtue V2(c)).

Apart from the issue of the objectivity of the specification of the prior, by and large the objectivist Bayesian approach has similar advantages and disadvantages regarding our list of virtues as its subjectivist cousin. Particularly it comes with the same difficulties regarding the issue of falsifiability from observations. Prior probabilities are connected to logical analysis of the situation rather than to betting rates for future observations as in de Finetti's subjectivist approach, which makes the connection of objectivist Bayesian prior probabilities to observations even weaker than in the subjectivist Bayesian approach (probabilistic logic has applications other than statistical data analysis, for which this may not be a problem).

The merit of objectivist Bayesianism is that the approach comes with a much stronger drive to justify prior distributions in a transparent way using principles that are as clear and general as possible.

5.5. Falsificationist Bayesianism, and frequentist probabilities in Bayesian statistics

For both subjectivist and objectivist Bayesians, probability models including both parameter priors and sampling models do not model the data-generating process, but rather represent

plausibility or belief from a certain point of view. Plausibility and belief models can be modified by data in ways that are specified *a priori*, but they cannot be falsified by data.

In much applied Bayesian work, in contrast, the sampling model is interpreted, explicitly or implicitly, as representing the data-generating process in a frequentist or similar way, and parameter priors and posteriors are interpreted as giving information about what is known about the 'true' parameter values. It has been argued that such work does not directly run counter to the subjectivist or objectivist philosophy, because the 'true parameter values' can often be interpreted as expected large sample functions given the prior model (Bernardo and Smith, 1994), but the way in which classical subjectivist or objectivist statistical data analysis is determined by the untestable prior assignments is seen as unsatisfactory by many statisticians.

In any case, the frequentist interpretation of a probability distribution as 'data generator' is regularly used to investigate how Bayesian analyses perform under such assumptions, theoretically, often by analysis of asymptotic properties or by simulation. Wasserman (2006) called Bayesian methods with good frequentist properties 'objective', referring to the 'representing things in the observer-independent world' sense of objectivity, but also providing a connection of Bayesian models to observables (virtue V4(a)). Rubin (1984) discussed frequentist approaches for studying the characteristics of Bayesian methods under misspecified models, i.e. stability (virtue V7).

The suggestion of testing aspects of the prior distribution by observations using error statistical techniques has been around for some time (Box, 1980). Gelman and Shalizi (2013) incorporated this in an outline of what we refer to here as 'falsificationist Bayesianism', a philosophy that openly deviates from both objectivist and subjectivist Bayesianism, integrating Bayesian methodology with an interpretation of probability that can be seen as frequentist in a wide sense and with an error statistical approach to testing assumptions in a bid to satisfy virtue V4(b).

Falsificationist Bayesianism follows the frequentist interpretation of the probabilities that is formalized by the sampling model given a true parameter, so that these models can be tested by using frequentist inference (with the limitations that such techniques have, as discussed in Section 5.2). Gelman and Shalizi (2013) argued, as some frequentists do, that such models are idealizations and should not be believed to be literally true, but that the scientific process proceeds from simplified models through test and potential falsification by improving the models where they are found to be deficient.

To put it another way, it is desirable for Bayesian intervals to have close to nominal coverage both conditionally on any observables and unconditionally; the desire for this coverage leads naturally to calibration checks, which in turn motivates the modification or even rejection of models that are not well calibrated empirically. This process serves the correspondence to observable reality (virtue V4) while putting more of a burden on transparency (virtue V1) and stability (virtue V7) in that the ultimate choice of model can depend on the decision of what aspects of the fitted model will be checked.

A key issue regarding transparency of falsificationist Bayesianism is how to interpret the parameter prior, which does not usually (if occasionally) refer to a real mechanism that produces frequencies. Major options are firstly to interpret the parameter prior in a frequentist way, as formalizing a more or less idealized data-generating process generating parameter values. A bold idealization would be to view 'all kinds of potential studies with the (statistically) same parameter' as the relevant population, even if the studies are about different topics with different variables, in which case more realizations exist, but it is difficult to view a specific study of interest as a 'random draw' from such a population.

Alternatively, the parameter prior may be seen as a purely technical device, serving aims such as regularization, without making any even idealized assumption that it corresponds to

anything that 'exists' in the real world. In this case the posterior distribution does not have a proper direct interpretation, but statistics such as the posterior mean or uncertainty intervals could be interpreted on the basis of their frequentist properties.

Overall, falsificationist Bayesianism combines the virtue of error statistical falsifiability with virtues V5 and V6 connected to subjectivity. However, the flexibility of the falsificationist Bayesian approach—its openly iterative and tentative nature—creates problems regarding clarity and unification.

6. Discussion

6.1. Implications for statistical theory and practice

At the level of discourse, we would like to move beyond a subjective *versus* objective shouting match. But our goals are larger than this. Gelman and Shalizi (2013) on the philosophy of Bayesian statistics sought not just to clear the air but also to provide philosophical and rhetorical space for Bayesians to feel free to check their models and for applied statisticians who were concerned about model fit to feel comfortable with a Bayesian approach. In the present paper, our goals are for scientists and statisticians to achieve more of the specific positive qualities into which we decompose objectivity and subjectivity in Section 3.4. At the present time, we feel that concerns about objectivity are obstructing researchers trying out different ideas and considering different sources of inputs to their model, whereas an ideology of subjectivity is limiting the degree to which researchers are justifying and understanding their model.

There is a tendency for hard-core believers in objectivity needlessly to avoid the use of valuable external information in their analyses, and for subjectivists, but also for statisticians who want to make their results seem strong and uncontroversial, to leave their assumptions unexamined. We hope that our new framing of transparency, consensus, avoidance of bias, reference to observable reality, multiple perspectives, dependence on context and aims, investigation of stability and honesty about the researcher's position and decisions will give researchers of all stripes the impetus and, indeed, permission, to integrate different sources of information in their analyses, to state their assumptions more clearly and to trace these assumptions backwards to past data that justify them and forwards to future data that can be used to validate them.

Also, we believe that the pressure to appear objective has led to confusion and even dishonesty regarding data coding and analysis decisions which cannot be motivated in supposedly objective ways; see van Loo and Romeijn (2015) for a discussion of this point in the context of psychiatric diagnosis. We prefer to encourage a culture in which it is acceptable to be open about the reasons for which decisions are made, which may at times be a mathematical convenience, or the aim of the study, rather than strong theory or hard data. It should be recognized openly that the aim of statistical modelling is not always to make the model as close as possible to observer-independent reality (which always requires idealization anyway), and that some decisions are made, for example, to make outcomes more easily interpretable for specific target audiences.

Our key points are as follows:

(a) multiple perspectives correspond to multiple lines of reasoning, not merely to mindless and unjustified guesses and
(b) what is needed is not just a prior distribution or a tuning parameter, but a statistical approach in which these choices can be grounded, either empirically or by connecting them in a transparent way to the context and aim of the analysis.

For these reasons, *we do not think it at all accurate to limit Bayesian inference to 'the analysis*

of subjective beliefs'. Yes, Bayesian analysis can be expressed in terms of subjective beliefs, but it can also be applied to other settings that have nothing to do with beliefs (except to the extent that all scientific inquiries are ultimately about what is believed about the world).

Similarly, *we would not limit classical statistical inference to 'the analysis of simple random samples'.* Classical methods of hypothesis testing, estimation, and data reduction can be applied to all sorts of problems that do not involve random sampling. There is no need to limit the applications of these methods to a narrow set of sampling or randomization problems; rather, it is important to clarify the foundation for using the mathematical models for a larger class of problems.

6.2. Beyond 'objective' and 'subjective'

The list in Table 1 in Section 3.4 is the core of the paper. The list may not be complete, and such a list may also be systematized in different ways. Particularly, we developed the list having particularly applied statistics in mind, and we may have missed aspects of objectivity and subjectivity that are not connected in some sense to statistics. In any case, we believe that the given list can be helpful in practice for researchers, for justifying and explaining their choices, and for recipients of research work, for checking to what extent the virtues listed are practised in scientific work. A key issue here is transparency, which is required for checking all the other virtues. Another key issue is that subjectivity in science is not something to be avoided at any cost, but that multiple perspectives and context dependence are actually basic conditions of scientific inquiry, which should be explicitly acknowledged and taken into account by researchers. We think that this is much more constructive than the simple objective–subjective duality.

We do not think that this advice represents empty truisms of the 'mom and apple pie' variety. In fact, we repeatedly encounter publications in top scientific journals that fall foul of these virtues, which indicates to us that the underlying principles are subtle.

Instead of pointing at specific bad examples, here is a list of some common problems (discussed, for example, in Gelman (2015) and Gelman and Zelizer (2015)), where we believe that exercising one or more of our listed virtues would improve matters:

(a) presenting analyses that are contingent on data without explaining the exploration and selection process and without even acknowledging that it took place;
(b) justifying decisions by reference to specific literature without acknowledging that what was cited may be controversial, not applicable in the given situation or without proper justification in the cited literature as well (or not justifying the decisions at all);
(c) failure to reflect on whether model assumptions are reasonable in the given situation, what effect it would have if they were violated or whether alternative models and approaches could be reasonable as well;
(d) choosing methods because they do not require tuning or are automatic and therefore seem 'objective' without discussing whether the methods chosen can handle the data more appropriately in the given situation than alternative methods with tuning;
(e) choosing methods for the main reason that they 'do not require assumptions' without realizing that every method is based on implicit assumptions about how to treat the data appropriately, regardless of whether these are stated in terms of statistical models;
(f) choosing Bayesian priors without justification or explanation of what they mean and imply;
(g) using non-standard methodology without justifying the deviation from standard approaches (where they exist);
(h) using standard approaches without discussion of their appropriateness in a specific context.

Most of these are concerned with the unwillingness to admit to having made decisions, to justify them, and to take into account alternative possible views that may be equally reasonable. In some sense perhaps this can be justified on the basis of a sociological model of the scientific process in which each paper presents just one view, and then the different perspectives battle it out. But we think that this idea ignores the importance of communication and facilitating consensus for science. Scientists normally believe that each analysis aims at the truth and, if different analyses give different results, this is not because there are different conflicting truths but rather because different analysts have different aims, perspectives and access to different information. Letting the issue aside of whether it makes sense to talk of the existence of different truths or not, we see aiming at general agreement in free exchange as essential to science and, the more perspectives that are taken into account, the more the scientific process is supported.

We see the listed virtues as ideals which in practice cannot generally be fully achieved in any real project. For example, tracing all assumptions to observations and making them checkable by observable data is impossible because one can always ask whether and why results from the specific observations that are used should generalize to other times and other situations. As mentioned in Section 5.1, ultimately a rationale for treating different situations as 'identical and independent' or 'exchangeable' needs to be constructed by human thought (people may appeal to historical successes for justifying such idealizations, but this does not help much regarding specific applications). At some point—but, we hope, not too early—researchers must resort to somewhat arbitrary choices that can be justified only by logic or convention, if that.

And it is likewise unrealistic to suppose that we can capture all the relevant perspectives on any scientific problem. Nonetheless, we believe that it is useful to set these as goals which, in contrast with the inherently opposed concepts of 'objectivity' and 'subjectivity', can be approached together.

Acknowledgements

We thank the US National Science Foundation and Office of Naval Research for partial support of this work, and Sebastian Weber, Jay Kadane, Arthur Dempster, Michael Betancourt, Michael Zyphur, E. J. Wagenmakers, Deborah Mayo, James Berger, Prasanta Bandyopadhyay, Laurie Paul, Jan-Willem Romeijn, Gianluca Baio, Keith O'Rourke, Laurie Davies and the reviewers for helpful comments.

Appendix A: Objectivity in the philosophy of science

Megill (1994) listed four basic senses of objectivity: 'absolute objectivity', in the sense of 'representing the things as they really are' (independently of an observer), 'disciplinary objectivity', referring to a consensus among experts within a discipline and highlighting the role of communication and negotiation, 'procedural objectivity', in the sense of following rules that are independent of the individual researcher, and 'dialectical objectivity', referring to active human 'objectification' required to make phenomena communicable and measurable so that they can then be treated in an objective way so that different subjects can understand them in the same way. These ideas appear under various names in many places in the literature. Porter (1996) listed the ideal of impartiality of observers as another sense of objectivity. Douglas (2004) distinguished three modes of objectivity: human interaction with the world, individual thought processes and processes to reach an agreement. Daston and Galison (2007) called the ideal of scientific images that attempt to capture reality in an unmanipulated way 'mechanical objectivity' as opposed to 'structural objectivity', which refers to mathematical and logical structures. The latter emerged from the insight of scientists and philosophers such as Helmholtz and Poincaré that observation of reality cannot exclude the observer and will never be as reliable and pure as 'mechanical objectivists' would hope.

More generally, virtually all senses of objectivity have been criticized at some point in history for being unachievable, which often prompted the postulation of new scientific virtues and new senses of objectivity.

For example, the realist ideal of 'absolute objectivity' has been branded as metaphysical, meaningless and illusory by positivists including Pearson (1911), and more contemporarily by empiricists such as van Fraassen (1980). The latter took observability and the ability of theory to account for observed facts as objective from an antirealist perspective.

Some writers even criticize the idea that objectivity is a generally desirable virtue in science, e.g. for its implication of a denial of the specific conditions of an observer's point of view (Feyerabend, 1978; MacKinnon, 1987; Maturana, 1988) and its use as a rhetorical device or tool of power (see Fuchs (1997) for a critical overview of such ideas).

The core benefit of such controversies around objectivity and subjectivity for statisticians is the elaboration of aspects of good science, which should inform statistical data analysis and decision making. Hacking (2015) wrote a paper called 'Let's not talk about objectivity', and with him we believe that, for discussing the practice of statistics (or more generally science), the objectivity *versus* subjectivity discourse should be replaced by looking at more specific virtues of scientific work, the awareness of which could have a more direct influence on the work of scientists. The virtues that we have listed in Section 3 are all connected either to senses of objectivity as summarized above, or to reasons for criticizing certain concepts of objectivity.

References

Agresti, A. and Coull, B. A. (1998) Approximate is better than exact for interval estimation of binomial proportions. *Am. Statistn*, **52**, 119–126.

Alpert, M. and Raiffa, H. (1984) A progress report on the training of probability assessors. In *Judgment Under Uncertainty: Heuristics and Biases* (eds D. Kahneman, P. Slovic and A. Tversky), pp. 294–305. New York: Cambridge University Press.

Bayarri, M. J., Berger, J. O., Forte, A. and García-Donato, G. (2012) Criteria for Bayesian model choice with application to variable selection. *Ann. Statist.*, **40**, 1550–1577.

Berger, J. O. (2006) The case for objective Bayesian analysis. *Baysn Anal.*, **1**, 385–402.

Berger, J. O., Bernardo, J. M. and Sun, D. (2009) The formal definition of reference priors. *Ann. Statist.*, **37**, 905–938.

Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013) Valid post-selection inference. *Ann. Statist.*, **41**, 802–837.

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory*. Chichester: Wiley.

Box, G. E. P. (1980) Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *J. R. Statist. Soc.* A, **143**, 383–430.

Box, G. E. P. (1983) An apology for ecumenism in statistics. In *Scientific Inference, Data Analysis, and Robustness* (eds G. E. P. Box, T. Leonard and C. F. Wu), pp. 51–84. New York: Academic Press.

Candler, J., Holder, H., Hosali, S., Payne, A. M., Tsang, T. and Vizard, P. (2011) Human rights measurement framework: prototype panels, indicator set and evidence base. *Research Report 81*. Equality and Human Rights Commission, Manchester.

Chang, H. (2012) *Is Water H_2O?: Evidence, Realism and Pluralism*. Dordrecht: Springer.

Cox, R. T. (1961) *The Algebra of Probable Inference*. Baltimore: Johns Hopkins University Press.

Cox, D. and Mayo, D. G. (2010) Objectivity and conditionality in frequentist inference. In *Error and Inference* (eds D. G. Mayo and A. Spanos), pp. 276–304. Cambridge: Cambridge University Press.

Daston, L. and Galison, P. (2007) *Objectivity*. New York: Zone Books.

Davies, P. L. (2014) *Data Analysis and Approximate Models*. Boca Raton: CRC Press.

Dawid, A. P. (1982a) Intersubjective statistical models. In *Exchangeability in Probability and Statistics* (eds G. Koch and F. Spizichino), pp. 217–232. Amsterdam: North-Holland.

Dawid, A. P. (1982b) The well-calibrated Bayesian. *J. Am. Statist. Ass.*, **77**, 605–610.

Desrosieres, A. (2002) *The Politics of Large Numbers*. Boston: Harvard University Press.

Dick, P. K. (1981) *VALIS*. New York: Bantam Books.

Douglas, H. (2004) The irreducible complexity of objectivity. *Synthese*, **138**, 453–473.

Douglas, H. (2009) *Science, Policy and the Value-free Ideal*. Pittsburgh: University of Pittsburgh Press.

Erev, I., Wallsten, T. S. and Budescu, D. V. (1994) Simultaneous over- and underconfidence: the role of error in judgment processes. *Psychol. Rev.*, **101**, 519–527.

Erikson, R. S., Panagopoulos, C. and Wlezien, C. (2004) Likely (and unlikely) voters and the assessment of campaign dynamics. *Publ. Opin. Q.*, **68**, 588–601.

Everitt, B. S., Landau, S., Leese, M. and Stahl, D. (2011) *Cluster Analysis*, 5th edn. Chichester: Wiley.

Feyerabend, P. (1978) *Science in a Free Society*. London: New Left Books.

Fine, T. L. (1973) *Theories of Probability*. Waltham: Academic Press.

de Finetti, B. (1974) *Theory of Probability*. New York: Wiley.

Fisher, R. (1955) Statistical methods and scientific induction. *J. R. Statist. Soc.* B, **17**, 69–78.

van Fraassen, B. (1980) *The Scientific Image*. Oxford: Oxford University Press.
Fuchs, S. (1997) A sociological theory of objectivity. *Sci. Stud.*, **11**, 4–26.
Gelman, A. (2008) The folk theorem of statistical computing. *Statistical modeling, causal inference, and social science blog*, May 13th. (Available from `http://andrewgelman.com/2008/05/13/the_folk_theore/`.)
Gelman, A. (2013) Whither the "bet on sparsity principle" in a nonsparse world? *Statistical modeling, causal inference, and social science blog*, Feb. 25th (Available from `http://andrewgelman.com/2013/12/16/whither-the-bet-on-sparsity-principle-in-a-nonsparse-world/`.)
Gelman, A. (2014a) How do we choose our default methods? In *Past, Present, and Future of Statistical Science* (eds X. Lin, C. Genest, D. L. Banks, G. Molenberghs, D. W. Scott and J. L. Wang), pp. 293–301. London: Chapman and Hall.
Gelman, A. (2014b) Basketball stats: don't model the probability of win, model the expected score differential. *Statistical modeling, causal inference, and social science blog*, Feb 25th. (Available from `http://andrewgelman.com/2014/02/25/basketball-stats-dont-model-probability-win-model-expected-score-differential/`.)
Gelman, A. (2014c) President of American Association of Buggy-Whip Manufacturers takes a strong stand against internal combustion engine, argues that the so-called "automobile" has "little grounding in theory" and that "results can vary widely based on the particular fuel that is used". *Statistical modeling, causal inference, and social science blog.* (Available from `http://andrewgelman.com/2014/08/06/president-american-association-buggy-whip-manufacturers-takes-strong-stand-internal-combustion-engine-argues-called-automobile-little-grounding-theory/`.)
Gelman, A. (2015) The connection between varying treatment effects and the crisis of unreplicable research: a Bayesian perspective. *J. Mangmnt*, **41**, 632–643.
Gelman, A. and Basbøll, T. (2013) To throw away data: plagiarism as a statistical crime. *Am. Scient.*, **101**, 168–171.
Gelman, A., Bois, F. Y. and Jiang, J. (1996) Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *J. Am. Statist. Ass.*, **91**, 1400–1412.
Gelman, A., and Carlin, J. B. (2014) Beyond power calculations: assessing Type S (sign) and Type M (magnitude) errors. *Perspect. Psychol. Sci.*, **9**, 641–651.
Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D., Vehtari, A. and Rubin, D. B. (2013) *Bayesian Data Analysis*, 3rd edn. London: Chapman and Hall.
Gelman, A., Goel, S., Rivers, D. and Rothschild, D. (2016) The mythical swing voter. *Q. J. Polit. Sci.*, **11**, 103–130.
Gelman, A. and Loken, E. (2014) The statistical crisis in science. *Am. Scient.*, **102**, 460–465.
Gelman, A. and O'Rourke, K. (2015) Convincing evidence. In *Roles, Trust, and Reputation in Social Media Knowledge Markets* (eds S. Matei and E. Bertino). New York: Springer.
Gelman, A. and Shalizi, C. (2013) Philosophy and the practice of Bayesian statistics (with discussion). *Br. J. Math. Statist. Psychol.*, **66**, 8–80.
Gelman, A. and Zelizer, A. (2015) Evidence on the deleterious impact of sustained use of polynomial regression on causal inference. *Res. Polit.*, **2**, 1–7.
Gillies, D. (2000) *Philosophical Theories of Probability*. London: Routledge.
Greenland, S. (2012) Transparency and disclosure, neutrality and balance: shared values or just shared words? *J. Epidem. Commty Hlth*, **66**, 967–970.
Hacking, I. (2015) Let's not talk about objectivity. In *Objectivity in Science* (eds F. Padovani, A. Richardson and J. Y. Tsou), pp. 19–33. Cham: Springer.
Hennig, C. (2010). Mathematical models and reality: a constructivist perspective. *Foundns Sci.*, **15**, 29–48.
Hennig, C. and Liao, T. F. (2013) How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification (with discussion). *Appl. Statist.*, **62**, 309–369.
Hennig, C. and Lin, C.-J. (2015) Flexible parametric bootstrap for testing homogeneity against clustering and assessing the number of clusters. *Statist. Comput.*, **25**, 821–833.
Huber, P. J. and Ronchetti, E. M. (2009) *Robust Statistics*, 2nd edn. New York: Wiley.
Jaynes, E. T. (2003) *Probability Theory: the Logic of Science*. Cambridge: Cambridge University Press.
Kahneman, D. (1999) Objective happiness. In *Well-being: Foundations of Hedonic Psychology*, pp. 3–25. New York: Russell Sage Foundation Press.
Kass, R. E. and Wasserman, L. (1996) The selection of prior distributions by formal rules. *J. Am. Statist. Ass.*, **91**, 1343–1370.
Kaufman, L.and Rousseeuw, P. J. (1990) *Finding Groups in Data*. New York: Wiley.
Kendall, M. G. (1949) On the reconciliation of theories of probability. *Biometrika*, **36**, 101–116.
Keynes, J. M. (1936) *The General Theory of Employment, Interest and Money*. London: Macmillan.
Knight, F. H. (1921) *Risk, Uncertainty, and Profit*. Boston: Hart, Schaffner and Marx.
Little, R. J. (2012) Calibrated Bayes, an alternative inferential paradigm for official statistics. *J. Off. Statist.*, **28**, 309–334.
van Loo, H. M. and Romeijn, J. W. (2015) Psychiatric comorbidity: fact or artifact? *Theoret. Med. Bioeth.*, **36**, 41–60.
MacKinnon, C. (1987) *Feminism Unmodified*. Boston: Harvard University Press.
Maturana, H. R. (1988) Reality: the search for objectivity or the quest for a compelling argument. *Ir. J. Psychol.*, **9**, 25–82.

Mayo, D. G. (1996) *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.
Mayo, D. G. and Spanos, A. (2010) The error-statistical philosophy. In *Error and Inference* (eds D. G. Mayo and A. Spanos), pp. 15–27. New York: Cambridge University Press.
Megill, A. (1994) Four senses of objectivity. In *Rethinking Objectivity* (ed. A. Megill), pp. 1–20. Durham: Duke University Press.
Merry, S. E. (2011) Measuring the world: indicators, human rights, and global governance. *Curr. Anthrop.*, **52**, suppl. 3, S83–S95.
von Mises, R. (1957) *Probability, Statistics and Truth*, 2nd edn. New York: Dover Publications.
Pearson, E. S. (1955) Statistical concepts in their relation to reality. *J. R. Statist. Soc.* B, **17**, 204–207.
Pearson, K. (1911) *The Grammar of Science*. New York: Cosimo.
Porter, T. M. (1996) *Trust in Numbers: the Pursuit of Objectivity in Science and Public Life*. Princeton: Princeton University Press.
Rubin, D. B. (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, **12**, 1151–1172.
Sheiner, L. B. (1984) The population approach to pharmacokinetic data analysis: rationale and standard data analysis methods. *Drug Metablsm Rev.*, **15**, 153–171.
Simmons, J., Nelson, L. and Simonsohn, U. (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allow presenting anything as significant. *Psychol. Sci.*, **22**, 1359–1366.
Steegen, S., Tuerlinckx, F., Gelman, A. and Vanpaemel, W. (2016) Increasing transparency through a multiverse analysis. *Perspect. Psychol. Sci.*, **11**, 702–712.
Tibshirani, R. J. (2014) In praise of sparsity and convexity. In *Past, Present, and Future of Statistical Science* (eds X. Lin, C. Genest, D. L. Banks, G. Molenberghs, D. W. Scott and J. L. Wang), pp. 505–513. London: Chapman and Hall.
Tukey, J. W. (1962) The future of data analysis. *Ann. Math. Statist.*, **33**, 1–67.
Tukey, J. W. (1977) *Exploratory Data Analysis*. Reading: Addison-Wesley.
Vermunt, J. K. and Magidson, J. (2016) *Technical Guide for Latent GOLD 5.1: Basic, Advanced, and Syntax*. Belmont: Statistical Innovations.
Wang, W., Rothschild, D., Goel, S. and Gelman, A. (2015) Forecasting elections with non-representative polls. *Int. J. Forecast.*, **31**, 980–991.
Wasserman, L. (2006) Frequentist Bayes is objective (comment on articles by Berger and by Goldstein). *Baysn Anal.*, **1**, 451–456.
Weinberger, D. (2009) Transparency is the new objectivity. *Everything is miscellaneous blog,* July 19th. (Available from `http://www.everythingismiscellaneous.com/2009/07/19/transparency-is-the-new-objectivity/`.)
Yong, E. (2012) Nobel laureate challenges psychologists to clean up their act. *Nat. News*, Oct. 3rd. (Available from `http://www.nature.com/news/nobel-laureate-challenges-psychologists-to-clean-up-their-act-1.11535`.)

Discussion on the paper by Gelman and Hennig

Philip Dawid (*University of Cambridge*)

There are two strands to this paper. On the one hand, it aims to defuse the ultimately fruitless opposition between 'subjective' and 'objective' positions on the meaning of statistical models of reality. On the other hand—and to my mind more importantly—it supplies, in Table 1, a checklist for thinking about the whole process of choosing, justifying and analysing a statistical model. This can be thought of as an initiating contribution to the important but neglected field of 'metastatistics'—principles that should govern the whole process of how we should think about conducting statistical modelling and analysis, in relation to important questions about the real world. Although all the virtues listed deserve discussion, here I shall comment only on V4. It is far from obvious how to contrast a probabilistic model with observed outcomes in the real world. I have long been an advocate of the 'falsificationist Bayes' position and have made numerous attempts, both philosophical and technical, to develop this approach (Dawid, 1982, 1984, 1985, 2004; Seillier-Moiseiwitsch *et al.*, l992; Seillier-Moiseiwitsch and Dawid, 1993). Many of the considerations are equally relevant to frequentist model checking.

What about metastatistical vices?—some are mentioned in Section 6.2 of the paper. Another vice that has greatly concerned me is 'reification', which is to regard all aspects of your mental model as if they had real counterparts in the world. In my own falsificationist philosophy, probability lives in the model; outcomes live in the world: it is misleading to regard probabilities (or 'data-generating processes') as real attributes of the world. Another egregious but common example occurs in causal analysis, where a directed graphical model is fitted to observational data, and the arrows in the graph are interpreted as having causal meaning. I have pontificated against this vice in Dawid (2010).

Stimulated by tonight's metastatistical paper, I now make some suggestions of my own for this enterprise. I shall phrase these as a collection of questions about things to think about.

(a) Who am I and what is my standpoint? Examples could be
  (i) a subject matter expert,
  (ii) an external consultant,
  (iii) a disinterested scientist,
  (iv) an interested party,
  (v) a stakeholder,
  (vi) an individual or
  (vii) a team.
  The answer to this question will affect, among other things, what prior information (and whose) I should take into account.
(b) What is my audience? Who might be affected by my findings? Examples could be
  (i) just me,
  (ii) people who share the same background assumptions as I do,
  (iii) people who will need more convincing or
  (iv) the world at large.
(c) Which aspects of my model do I care about, and which can I afford to get wrong?
(d) Is this exercise a one-off, or part of a continuing cycle? Where and how does it fit into the overall scientific enterprise?
(e) What is my purpose in selecting and fitting a statistical model? Possible answers might be
  (i) prediction of new observations (single, batch or sequential),
  (ii) explanation of old observations,
  (iii) discovery of novel phenomena,
  (iv) understanding of underlying processes,
  (v) discovery of causal relationships or
  (vi) a guide to future actions.
  For example, there has been much heated argument about whether human activity has been responsible for global warming. But the real question is: what (if anything) can we now do to mitigate its effects? Although these questions are related, they are not identical.
(f) How shall I judge whether and when I have achieved my purpose?
(g) What kind of model is appropriate?—
  (i) a substantive model, accounting for, for example, physical understandings of the processes under investigation;
  (ii) a black box model, aiming to capture statistical regularities;
  (iii) a causal model, giving guidance on the effects of interventions.
  An interesting discussion of such points may be found in Christie *et al.* (2011), which is the output of an interdisciplinary project that brought together experts from a variety of applied and theoretical fields to compare their approaches to modelling. For example, when should a model incorporate complex differential equations standard in the field of study, and when is it more appropriate to fit a simple multiple regression?
(h) Why might I (possibly) want to fit a simple model?—
  (i) because I believe the underlying reality is (close to being) simple;
  (ii) because, without believing that, I think that a simpler model will be more stable to fit and will make better predictions.
  As an example, the currently popular field of sparsity modelling can be considered from either point of view. Which is appropriate may affect how the enterprise should be conducted.

Finally, if we are to take the metastatistical enterprise seriously, what meta-metastatistical principles might be relevant to it? How can we organize or prioritize metastatistical virtues and vices? Indeed, to what extent is it worth even trying to formalize this enterprise?—or should we just 'fly by the seat of our pants'?

I have greatly enjoyed this stimulating paper, and it gives me much pleasure to propose a vote of thanks to Gelman and Hennig.

Christian P. Robert (*Université Paris Dauphine, University of Warwick, Coventry, and Centre de Recherche en Economie et Statistique, Paris*)

Whereas I fully agree with Gelman and Hennig's perspective that there are more objectivity issues in

statistics than the mere choice of a prior distribution in Bayesian statistics, I first doubt switching terms as proposed therein will clarify the subjective nature of the game for everyday users and second feel that there are deeper issues with the foundations of statistics that stand beyond redemption. Although surprised at seeing a paper entirely devoted to (necessarily subjective) philosophical opinions about statistics I obviously welcome the opportunity for such a discussion.

Indeed, 'statistics cannot do without data' but the paper does not really broach the question whether or not statistics cannot do *without probability*. Although this may sound like a *lieu commun*, let us recall that a statistical analysis almost invariably starts with the premise that the data are *random*. However, the very notion of randomness is quite elusive; hence this aspect truly fits within the paper's topic—without even mentioning the impossibility of establishing randomness for a given phenomenon, barring maybe instrumental error. This query extends to the notion of a probabilistic generative model and it relates more directly to the repeatability assumption that should not be taken for granted in most realistic situations.

The central message of the paper is that statistical analyses should be open about the many choices made in selecting an estimate or a decision and about the forking paths of alternative resolutions. Given that very few categories of statisticians take pride in their subjectivity, but rather use this term as derogatory for other categories, I fear that the proposal stands little chance of seeing this primacy of objectivity claims resolved, even though I agree that

(a) we should move beyond a distinction that does not reflect the complexity and richness of statistical practice and
(b) we should embrace and promote uncertainty, diversity and relativity in our statistical analysis.

As the discussions in Sections 2 and 5 make clear, all statistical procedures involve subjective or operator-dependent choices and calibration, either plainly acknowledged or hidden under the carpet. This is why I would add (at least) two points to the virtues of subjectivity to Section 3.4 that is central to the paper's message:

(a) spelling out unverifiable assumptions about data collection;
(b) awareness of calibration of tuning parameters;

though I do not see consensus (item (b)) as a necessary virtue.

In fact, when going through the examination of objectivity claims by the major threads of formalized statistical analysis, I have the feeling of exploring many small worlds (in Lindley's words) rather than the entire spectrum of statistical methodologies. For instance, frequentism seems to be reduced to asymptotics, while completely missing the area of non-parametrics. (The latter should not be considered to be 'more' objective, but it offers the advantage of loosening model specification.) Frequentist inference is mostly equated with the error statistical proposal of Mayo (1996), despite the availability of other and more mainstream perspectives. In particular, except for the reference to Davies (2014), the M-open view seems to be missing from the picture, despite attempting to provide reasoning '*outside the box*'. From a Bayesian perspective, the discussions of subjective, objective and falsificationist—missing empirical—Bayes do not really add to the debate between these three branches, apart from suggesting that we should give up such value-loaded categories. I came to agree mostly with the subjectivist approach on the ground of relativity, in that the outcome is always relative to a well-specified universe and that it can solely be measured in terms of that reference. I further find the characterization of the objectivist branch somehow restrictive, by focusing solely on Jaynes's (2003) maxent solution (which itself depends on many subjective choices). Hence, this section is missing in the corpus of work about creating priors with guaranteed frequentist or asymptotic properties. Furthermore, it operates under the impression that an objective Bayes analysis should always achieve the *same conclusion*, which misses the point of an automated derivation of a reference prior construct. That many automatizations are feasible and worth advocating nicely fits with the above relativity principle. I also find the defence of the falsificationist perspective, i.e. of essentially Gelman and Shalizi (2013), both much less critical and extensive, in that, again, this is not what one could call a standard approach to statistics.

In conclusion, the paper is appealing in calling for an end to the 'objectivier than thou' argument, but more difficult to perceive as launching a move towards a change in statistical practice. On the positive side, it exposes the need to spell out the inputs—from an operator—leading to a statistical analysis, both for replicability or for reproducibility reasons and for 'objectivity' purposes, although solely conscious perceived choices can be uncovered this way. It also reinforces the call for model awareness, by which I mean a critical stance on *all* modelling inputs, including priors, i.e. a disbelief that any model is true, applying to statistical procedures Popper's critical rationalism. This has major consequences for Bayesian modelling in that, as advocated in Gelman and Shalizi (2013), and Evans (2015), sampling and prior models

should be given the opportunity to be updated when inappropriate for the data at hand. A potential if unrealistic outcome of this paper would be to impose not only the production of all conscious choices made in the construction process, but also through the posting of (true or pseudo) data and of relevant code for all publications involving a statistical analysis. On the negative side, the proposal is far too idealistic in that most users (and most makers) of statistics cannot or would not spell out their assumptions and choices, being unaware of or unapologetic about these. This can be seen as a central difficulty with statistics as a service discipline, namely that almost anyone anywhere can produce an estimate or a p-value without ever being proven wrong. It is therefore difficult to fathom how the epistemological argument therein—that objective *versus* subjective is a meaningless opposition—could profit statistical methodology, even assuming that the list in Section 3.4 is made compulsory. The eight sins listed in the final section would require statistics expert reviewers for all publications, whereas almost never do journals outside our field call for statistics experts within referees. Apart from banning all statistics arguments from journals, I am afraid that there is no hope for a major improvement in that corner.

It is thus my great pleasure to second the vote of thanks for this multifaceted paper that helps to strengthen the foundations of our field.

The vote of thanks was passed by acclamation.

Richard D. Morey (*University of Cardiff*)

Many thanks go to Gelman and Hennig for their engaging discussion. As an applied statistician, I find the philosophical foundations of statistics fascinating and perplexing: fascinating, because they are inquiry into learning from observations, which is a topic central to science; perplexing, because central concepts sometimes appear trivial under examination (e.g. frequentist confidence or Bayesian coherence).

In the body of statistics, foundations are the supporting skeletal frame. Foundations can seem trivial because they are incapable of 'heavy lifting', collapsing without the connective tissue of pragmatics. Gelman and Hennig do not so much replace foundational ideas like objectivity and subjectivity as metaphorically put flesh on their bones, connecting them in various ways that enable them to support real work.

I briefly comment on the relationship between stability, consensus, correspondence to objective reality, and then parsimony.

Although stability, consensus and correspondence to objective reality are described as statistical virtues, they are best subsumed under the virtue *selective sensitivity*. We do not aim at stability or consensus *per se*; we aim at methods that allow us to choose to be sensitive to variance in *this*, but not in *that* (or we would prefer variance in some opinions but not others, according to the evidence). Although sensitivity is discussed, it is in the context of avoiding sensitivity to arbitrary assumptions. But of course we do not want stable inferences even when a parameter of interest is varied; we want sensitivity to that variance. Selective sensitivity, not stability, is the underlying virtue.

Relatedly, correspondence to objective reality is not a virtue of a method—not in the same way as stability or sensitivity. Sensitivity is the important methodological virtue, connecting a method to what might occur in *hypothetical* realities and ensuring that our inference is reliable in this reality. Joint consideration of selectivity and sensitivity makes clear that stability has a cost: increasing selectivity (e.g. robustness) may decrease sensitivity (e.g. power) or interpretability.

Finally, I missed discussion of parsimony. An informal preference for simplicity is essential to data analysis, from statistical regularization all the way up to the choice of model or technique to clear graphical representations. Although I know Gelman has questioned the usefulness of formal statistical parsimony, in practice continuous model expansion seems to favour an informal notion of parsimony in deciding when to stop expanding. Perhaps parsimony and transparency go hand in hand: simplicity is a means of ensuring that we understand our models, even when they are in some sense 'approximations'.

Marc Vandemeulebroecke (*Novartis, Basel*)

I thank Gelman and Hennig for giving me the opportunity to comment on this highly interesting paper. It is rich in content and carries an important message: in statistics (and more generally science), let us not avoid good reason and judgement—let us avoid labels. As we can only access reality through observations, multiple perspectives are unavoidable, and scientific progress can only happen through a process of interaction. So let us be open about multiple perspectives and sources of information, honest about our assumptions and goals, and specific in our language. In particular, let us replace the ambiguous and loaded labels 'subjective' and 'objective' by the seven virtues listed in Table 1 of the paper.

It is interesting that the terms 'subject(ive)' and 'object(ive)' not only carry various meanings and connotations today—they have also changed their meaning significantly over time. Originally, they were Latin translations of the Greek (notably Aristotle's (Ackrill, 1963)) terms '*hypokeimenon*' and '*antikeimenon*', literally 'what is lying underneath' (the true underlying essence) and 'what is lying against' (the opposite or target). Still in mediaeval philosophy, the '*esse subiectivum*' meant the real (observer-independent) state of being, whereas the '*esse obiectivum*' was what is observed, intended or thought by someone. It is only with Baumgarten and Kant that these terms receive approximately their modern meanings. For Kant (Gregor, 2015), '*subjektiv*' belongs to or comes from an individual, whereas '*objektiv*' is valid for every rational being.

I close with two thoughts from the perspective of a statistician working in drug development. First, in my view, the authors focus primarily on situations where we want to *learn* something about the world (the voter, the drug . . .). However, in drug development, at some point we must *confirm* what we have learned (Sheiner, 1997). Here, emphasis is placed on type I error control and prespecification of analyses. Some of the practical implications of the paper, such as being open to revising assumptions, priors or the model in light of the data, may need modification in this setting. Second, replacing two notions by seven can be much to ask from everyday statistical practitioners. It reminds me of the debate about dichotomized end points and 'responder analyses' (e.g. Senn (2003)). Also here, we are lured into a duality that appears simple but often hides many arbitrary underlying decisions. Still, its (perceived!) simplicity is so appealing that it continues to be widely used. We should try our best to remain alert with respect to the call by Gelman and Hennig.

Tom King (*Newcastle-upon-Tyne*)
Gelman and Hennig's focus on the practice of statistics gives the impression of inferences produced in splendid isolation, but the purpose of inference is to influence thinking. Of course, this means presenting analysis to other people, and making choices about how to do so, and one of the common ways we do this is by using graphs. The authors do mention visualization, in the process of analysis, but do not follow this through to communicating analysis to others. In a purely objective frame, graphs would not be necessary as the tabulated estimates covered all of the information. Indeed their history is that graphs were rarely used in the 19th century, with tables preferred for reasons of printing cost, and justified by simpler methods. But, since they were first widely used to reach illiterate audiences, graphs have become common, and recommended, inclusions in reports of statistical analyses.

Graphs allow the accurate perception of some patterns and proportions, while having regard to sequence and some correlation, which is not immediate in tables. So, given these cognitive limitations of people, Tufte (1983) developed principles which can be read as espousing objectivity of presentation. Although including distortions is a concern, other principles like an abhorrence of 'chart junk' and the maximization of the 'data:ink ratio' are design choices. Visual information theory comfortably supports a frame of descriptive semantics, the lack of distortion above as pragmatics, and a persuasive rhetoric. Abstracting the particular presentation is the analyst's subjective choice that there is a story to show which might otherwise be missed. Tufte's (1983) principles focus on presenting this semantic unembellished, objectively as it were, but this pragmatism neglects the rhetoric aspect.

An example of where this objectivity falls down is in the portability of an image, that it can be presented to others, independently of the analyst, and a rhetorical narrative introduced. Feinstein's (2003) graph attracted some notoriety because measurement error led to exaggerated presentation of the social prospects of children's development. Specifically, a feature of the graph, two lines crossing, was identified as persuasive, although it was not the focus of inference. In sum, if we hope that an analysis will be used to persuade, and include a graphic to support this, we should not expect our interpretation to accompany it. Moreover, if we do not supply any rhetoric at all, we should not expect our interlocutors to go without, despite their possibly flawed interpretation.

Henry Wynn (*London School of Economics and Political Science*)
The objective–subjective conundrum goes deep into philosophy. Many think that Immanuel Kant got it about right. We seem to be equipped with some pure intuitions which allow us to award structure to reality. He put space, time and causation on his list and maybe we should add energy and probability itself. But there is a barrier preventing us from gaining access to the hidden 'things as they are'. Schopenhauer more or less agreed but thought that we had privileged access via our own bodies and minds. This is backed up by recent work on the brain showing that subconscious intention seems to start before the conscious mind kicks in. Maybe belief is buried even deeper, driving intention.

The Bayesian approach has problems when the prior distribution is very 'informative' but is then radically contradicted by the data. Suppose that I am almost certain that my keys are on my desk and have uniform

mass of prior probability on a finite number of other locations then (under simple conditions) the optimal Bayesian experiment is to look on my desk. But then if my keys are not on my desk I am left with a uniform posterior on those other places. The Shannon information may go down from prior to posterior; I am surprised.

There has been a rush towards objective Bayesianism (Jeffreys priors and the like) with coverage and consistency properties similar to frequentist methods. This is good diplomacy but really misses the point. We urgently need ways to combine expert judgement with data and either the Bayesian method will remain the best or we need to spend more effort, not less, on the objective–subjective conundrum. In this sense this paper is a valuable contribution.

We can agree with Gelman and Hennig that any solution needs what the philosopher John Dewey calls 'reflection'. If the keys are not on my desk, I should think where I last saw them, try to remember my movements yesterday evening etc. Lazy is 'if my keys are not on my desk I don't know where they are'. Better is 'if my keys are not on my desk I had better think again'. This reflection, also called 'deliberation', is best carried out at a community level: a community of scientists or a social community. Lone belief and utility may cease to be as fashionable as they have been.

Kevin McConway (*The Open University, Milton Keynes*)

I welcome this paper because it moves discussion of what we do as statisticians, and why, from stereotyping labels of approaches mixed with sometimes inaccessible (albeit important) technical issues, to something that makes us think about what the point really is.

The first virtues on Gelman and Hennig's list, on transparency, imply that we should be open in communicating what we do, including the extent to which we demonstrate the other virtues. In my roles as statistical adviser to biomedical journals, and as explainer of statistical aspects of research papers to journalists (for the UK's Science Media Centre), I see many scientific papers with the conventional structure: introduction citing previous work, methods, results and discussion relating to the wider field and pointing out strengths and limitations. It strikes me that this structure, artificial though it can seem, does force authors to consider and communicate many aspects of virtue (on the understanding that these virtues apply much more widely than to statistical work alone). The introduction and methods sections will define concepts and explain plans and protocols (virtues V1(a) and V1(b)). Seeking consensus (V2) and impartiality (V3) should be covered in the introduction and discussion. A good methods section should deal with connection to observables, and with conditions for testing and falsification (V4), with the details demonstrated in the results section. The presence of virtues V5 and V6 can be judged from a decent introduction and discussion, even if the acknowledgement of the researchers' positions and fair consideration of other perspectives may need to be forced in by peer reviewers. Finally, in a study using quantitative evidence at least, there will usually be some analysis of sensitivity that relates to virtue V7. Thus a combination of conventional structure, journals' requirements on format of submissions and the demands of peer review do lead to publications where one can judge how virtuous the authors were. Ironically, in my experience the parts of scientific papers that tend to be least virtuous are those involving statistical analysis, which may not go beyond following generally accepted rules (virtue V2(b)) whether or not they are reasonable.

Papers on statistical methodology, for very good reasons, usually do not follow the same tightly stereotyped structures. That implies that statisticians may have to think harder about how to demonstrate their virtuousness than do, say, biomedical scientists. But that extra thought may itself be a good thing.

Robert Grant (*St George's University of London*)

Listening to this stimulating discussion, I began thinking about the dual activities of statistical inference and explanatory inference that we all undertake. Subjectivity and objectivity are at work here also, in rather simplistic forms, and could benefit from expansion into the terms that Gelman and Hennig propose. To clarify my terminology, consider the common problem of confounding. We can find a statistical association between, for example, maternal smoking and Down syndrome, as Chen *et al.* (1999) did (odds ratio 0.80; 95% confidence interval 0.65–0.98). The statistical inference is simply about the point estimate and its uncertainty. But then there is an explanatory inference that gives us a clue that this is misleading for causal inference: the smokers are younger and it is age that drives the observed difference. This is arrived at by intuition and subject expertise; confounding cannot be statistically detected. It feeds back into and refines the statistical inference because we are usually interested in the explanation (understanding), not the statistics (quantification). But there are problems here: explanations cannot be systematically sought or quantitatively compared; they are subject to many biases and can inflate error rates by using up what has been called 'researcher degrees of freedom'.

Explanation is very intuitive and is not prespecified. As a medical statistician, my work generally involves analysis of someone else's project (statistical inference), then discussion of the findings with the subject experts (explanatory inference) and then often some revision of the analysis. I expect the experts to do the explanatory part and the statistical part is then only as strong as their judgements. Explanation can be done with a subjective gloss ('that feels about right') or objective ('that matches what Smith and Jones published on the subject'), but the process of arriving at the explanation remains obscure. The philosopher Peter Lipton suggested that we gravitate towards explanations that are likely (that fit the data in some way) and lovely (that provide the most understanding, especially explaining other phenomena with minimal complexity) (Lipton, 2004). Gelman and Hennig's new terms are helpful but need to be applied to explanatory inferences also.

The following contributions were received in writing after the meeting.

Prasanta S. Bandyopadhyay (*Montana State University, Bozeman*)

I congratulate Andrew Gelman and Christian Hennig for a timely paper on the subjectivity–objectivity issues in statistics (Gelman and Shalizi (2013); see also Bandyopadhyay *et al.* (1996, 2014, 2015, 2016, 2017), Bandyopadhyay (2007), Bandyopadhyay and Brittan (2002), Bandyopadhyay and Boik (1999) and Bandyopadhyay and Forster (2011)). I shall however, raise some questions about their paper.

(a) Will the subjectivity–objectivity issue in statistics disappear if we use different expressions instead? We can presumably better understand it by making those terms clearer in terms of 'transparency', 'multiple perspectives', 'awareness of contextual dependence', etc. The authors attempt to 'decompose' subjectivity–objectivity to describe their nuances better. However, decomposing them in a nuanced way does not entail that the subjective–objective issues themselves are thereby solved. I sympathize with the idea of decomposition, but not necessarily with the paper's title.

(b) Some of the virtues they provided could be further fine grained, e.g. 'the awareness of contextual dependence' without branding it necessarily subjective. Influenced by Royall (1997), assuming a person's X-ray to be positive, I shall provide a scenario where one could ask at least *four types* of question concerning data. Here H1 is the hypothesis that the person has tuberculosis, and H2 is its denial (Table 2). Depending on which question one addresses, one could be a subjectivist. To address the evidence question, one need not be a subjectivist.

(c) The authors advocate 'falsificationist Bayesianism' as a novel account that goes beyond subjectivity–objectivity in statistics where one can *reject* a model and *replace* it by a better one. This gradual process of pruning models by better ones, *per* them, will lead to better science. The idea of rejection, however, goes against the idea of coherence, which is central to Bayesianism. So, falsificationist Bayesianism does not make sense if Bayesians give up the idea of coherence.

(d) Contrary to them, I advocate a combination of both subjective and objective features of inference within a robust Bayesian framework. To address the confirmation question in Table 2, I am a subjective Bayesian, whereas, to deal with the evidence question, I borrow ideas from a likelihoodist while being consistent with Bayesianism. To handle the prediction question, which involves model selection issues, one could use the Akaike information criterion (Akaike, 1973). However, as a Bayesian, I use the posterior predictive distribution where priors on models are constrained by simplicity constraints forcing me to be not a pure subjectivist But, one should be a subjective

Table 2. Four questions and statistical paradigms

Nature of the question	*Questions themselves*	*Statistical paradigms*
Confirmation question	Given the data, what should we *believe* and to what degree?	Subjective Bayesians
Evidence question	What do the data say about *evidence* for one hypothesis against its alternative and how much?	Likelihoodists or evidentialists
Prediction question	Given the data what is the *predictive accuracy* of the hypothesis?	Evidentialists or Bayesians of some version
Decision question	What should I *do*?	Subjective Bayesians

Bayesian to address the decision question. Thus, relevant questions contextualize subjectivity and objectivity without escaping them.

Richard J. Barker (*University of Otago, Dunedin*)

Gelman and Hennig are to be commended for their thoughtful critique of the notions of objectivity and subjectivity in statistical analysis. I am in full agreement with the underlying theme: that the notions of *objectivity* and *subjectivity* represent a misleading dichotomy—all statistical models require choices that cannot be guided by the data alone. The notion of transparency seems critical to me. Readers should always be given the opportunity to reach their own conclusions based on the data at hand and given their own choices.

I am not sure how this fits with their list of virtues but I see scientific *scepticism* as critical: we should question all analyses, especially our own. The statement 'dissent is essential to scientific progress' is an acknowledgement of the importance of scepticism. There is an understandable tendency to be overly admiring of our models, especially those that have complicated hierarchical structure. However, complexity in models is something to be suspicious of, especially if that structure is needed to account for sampling complexities or because of the inadequacies of the data.

As an applied statistician who uses Bayesian inference extensively, I am a heavy user of full probability models. Regardless of whether using Bayesian or frequentist inference I think of my probability models as analogies. I do not believe that they exactly depict reality. Rather I see them as convenient mechanisms that allow us to extrapolate from the abstract world of mathematics to the real world from which our data come. An axiomatic approach to probability-based reasoning is an appealing and powerful way to communicate uncertainty. But how convincing these descriptions are depends on the validity of the assumptions. It is here that the virtues seem relevant in that they allow readers to form their own views by helping to answer three fundamental questions that bear on the assumptions and the information we are treating as known.

(a) What are the bases for the assumptions used?
(b) What changes when we make different assumptions?
(c) What information are we ignoring?

Answering the last question can be especially difficult as it is typically unstated and often unjustified. In the end, most statistical sins turn out to be sins of conditioning.

David Bartholomew (*Sudbury*)

Gelman and Hennig are to be congratulated on their attempt to lift us out of the ruts of our inferential habits and to view statistical procedures from a more detached standpoint. I applaud their view that statistical methods fall squarely within the ambit of the scientific method. The true purpose of statistics is to learn something about the real world from data. Statistical methods are not a branch of pure mathematics and though reliance on elegance may build up credit with one's mathematical colleagues it may actually introduce spurious precision. 'Sufficiency', for example, is an important concept which is often too limited for situations in which it is used.

A good case could be made for structuring the authors' proposal on somewhat different lines. It is hazardous, of course, to try to match the authors' philosophical subtlety within the confines of a brief contribution such as this and I apologize in advance for the crudeness of my attempt. Following conventional bipolar thinking, we are invited to consider a spectrum of positions between 'subjective' at one extreme to 'objective' at the other. I would prefer to abandon this implied dichotomy and regard all statistical inference in a space of 'subjectivism' with 'objective' being a boundary point representing a distant, if unattainable, goal. Points in this subjective space might then be appropriately described in language like that set out in Table 1.

In this connection it is readily recognized that all so-called frequentist inferences involve a degree of subjectivity. For example, although the sample size may be treated as fixed, this may not actually be so. The choice may, in fact, have resulted from the resolution of conflict between competing interests, or the data may actually be the outcome of a sequential experiment with ill-defined and often unrecognized stopping rules. Conditioning on the sample size we ignore some information which may be relevant. Such 'objective' inferences may thus easily conceal an unacknowledged subjective input.

To change direction: very little attention is given by theoretical statisticians to the question of inference from finite populations. Yet this might serve as a prototype for all statistical inference. A finite population is a well-defined entity as is the notion of random sampling from it. Does this provide, I wonder, a model for inference—after introducing some suitable informal limiting process perhaps? Does this fit anywhere in the authors' scheme of thinking?

Michael Betancourt (*Columbia University, New York*)

I congratulate Gelman and Hennig on a thought-provoking paper, and a critical contribution to the ongoing development of transparent and reproducible scientific research. Only by moving beyond the crude portrayal of statistical methodologies as *objective* or *subjective* will we be able to evaluate fairly the established orthodoxies, and the factions within them, in the context of modern applications. And only with that evaluation will we be able to evolve applied statistics towards the frontiers of science and industry.

Because of the importance of this evolution, the details of the paper will no doubt spark a wealth of welcome discussion. For example, whereas I personally agree that transparency, correspondence to observational reality and stability are essential and largely independent virtues, I am hesitant about consensus, impartiality, multiple perspectives and context dependence. In particular, I do not see how consensus and impartiality can be separated from the acknowledgement of multiple perspectives and context dependence.

The history from which any consensus or impartiality is derived must be aggregated from multiple perspectives and contexts, and often those contexts are opaque and susceptible to misunderstanding. Consequently such derivations must themselves be evaluated as carefully as the assumptions fresh to a new analysis. Perhaps instead we should embrace statistics not as a methodology for resolving disagreements in science but rather as a methodology for communicating and discussing them. Motivated by the authors' own words, 'Consensus cannot be enforced; as a virtue it refers to behaviour that facilitates consensus', might we replace all four of these virtues with a single virtue that the ideal analysis acknowledges and considers, but not necessarily defers to, as much external information as possible?

Anne-Laure Boulesteix (*Ludwig-Maximilians University, Munich*) **and Ulrich Strasser** (*University of Innsbruck*)

Multiple facets of the multiplicity of perspectives

Gelman and Hennig's brilliant paper addresses crucial issues of utmost importance for the statistical community—and beyond. In light of their framework we revisit concepts from existing and related (statistical) works and draw parallels to other disciplines, both of which support the relevance of the described 'virtues'.

Neutral comparison studies (NCSs) (Boulesteix *et al.*, 2013, 2015) systematically evaluating methods using many (simulated and real) data sets produce evidence that contributes to generate *consensus* between researchers on the appropriateness of methods in various situations based on *observed reality* and that provides a basis for *transparent* decisions. NCSs exercise and encourage *impartiality* by assessing several methods without favouring a particular one; see related thoughts elaborated by the 'Strengthening analytical thinking for observational studies' initiative aimed at producing guidance documents for medical statistics (Sauerbrei *et al.*, 2014). Nevertheless, the *multiplicity of perspectives* will remain, not least because decisions regarding statistical analyses are *context dependent* and these complex dependences cannot be formalized and addressed within NCSs.

Related to the *multiplicity of perspectives*, serious problems may occur if researchers choose their analysis approach (only) because they yield the results they expected or hoped for (Boulesteix *et al.*, 2017), e.g. a significant effect of their 'favourite' variable (a form of *partiality* sometimes denoted as 'fishing for significance'). Considering the variability of the results across perspectives, statistical analyses would then be—as an extreme—a marketing strategy to promote ideas developed previously by researchers, independent of their data.

Regarding the *investigation of stability* and *awareness of the multiplicity of perspectives*, analysis and reporting practices differ across scientific fields. In some disciplines, such as health sciences, the *evaluation of the consequences of alternative decisions and assumptions*, termed 'sensitivity analyses', is often devoted little attention. In other scientific fields, an ensemble approach (using an entire set of credible models with the same input data to produce a range of outcomes) has proven to be a generally accepted methodology in providing an uncertainty estimate along with actual results, e.g. in climate modelling (Fig. 1). This 'ensembles approach' is even advocated on a higher level of implication, namely in the context of the 'grand challenges' for mankind in Earth system science for global sustainability (Reid *et al.*, 2010). They additionally stress the necessity of interdisciplinary research to encompass the intersection of global environmental change and sustainable development—another dimension of the *multiplicity of perspectives*?

Gilles Celeux (*Inria, Paris*)

I congratulate Andrew Gelman and Christian Hennig for their highly interesting and stimulating paper. I agree with their proposition to bring to the fore the attribute *transparency* instead of the attribute *objectivity*.

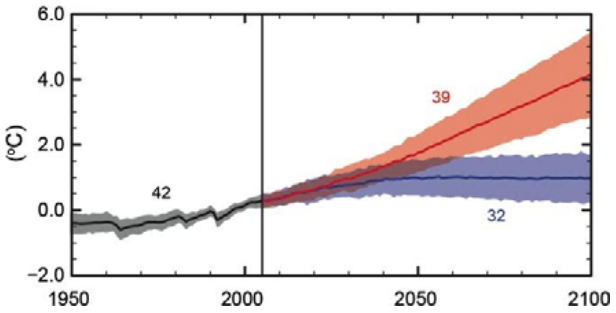

Fig. 1. Multimodel simulated change in global annual mean surface temperature from 1950 to 2100, given as temperature deviation from the mean of the period 1986–2005: as a measure of uncertainty of the simulation results, their range (minimum to maximum) is shown as shaded areas for low (■) and high (■) greenhouse gas emission scenarios; the dark area (■) is the modelled historical evolution; the numbers are the numbers of models used to calculate the respective multimodel means displayed as curves (simplified after Intergovernmental Panel on Climate Change (2013))

As a matter of fact, statistical models are not expected to explain or describe the world, but they can rather be expected to provide tools to act on it. For this very reason, transparency is desirable.

But, I am not sure that transparency is easy to ensure in the Bayesian framework with complex models. Actually, the influence of hyperparameters could be quite difficult to analyse in an informative setting and this task could appear to be even more difficult in a non-informative setting. In other words, in some circumstances, choosing prior distributions could appear to be a sorcerer's apprentice game that is hardly compatible with transparency. Anyhow, transparency of a Bayesian statistical analysis requires detailed (and expensive) sensitivity analyses as soon as the statistical models become somewhat complex.

Nicholas J. Cox (*University of Durham*)

I congratulate Gelman and Hennig on a lucid and provocative discussion.

The words *objectivity* and *subjectivity* can be broadly useful so long as they are qualified. The authors' checklist of terms should encourage finer distinctions and more nuanced debates. Most problems arise from exaggerated claims to objectivity, which are often unhelpful within the statistical sciences, and more likely to provoke critical or hostile reactions beyond them.

The analysis of the paper changes scale intermittently. In particular projects an individual or a team of researchers should try to be open about their subjective biases (theoretical, methodological, technical, etc.). In contrast fields or disciplines (say, voting studies or political science) are collective arenas in which debate leads unevenly towards a more objective consensus. Focus on which scale is being discussed is helpful: how far does subjectivity at *project* scale, long term, promote objectivity at *collective* scale by widening the space of ideas explored? Mavericks define the mainstream as well as conversely.

The authors' stance seems entirely compatible with statements from several earlier scholars. Jeffreys (1961) saw objectivity as a goal, not an achievement:

> 'if objectivity has any meaning at all, our aim must be *to find out* what is objective by means of observations'

(page 11; cf. pages 56 and 405). The philosophy of Karl Popper (e.g. Popper (1979)) bears the same message at greater length. Tukey's (1977), page 215, homespun wisdom may just summarize the problem but is still helpful:

> 'A great advantage of judgment is its ability to meet different needs. (The corresponding disadvantage is shown by the arguments among those who want to make different judgments.)'

Trimmed means may seem at best awkward and arbitrary. What could be good objective reasons for trimming 5% or 25% or any other fraction? Sensible seeming results are encouraging but not a rationale. One way to avoid the problem is to use all possible trimming amounts to provide a distribution signature (Rosenberger and Gasko, 1983; Davison and Hinkley, 1997; Cox, 2013).

Attitudes to transformation mix subjectivity and objectivity, or judgement and evidence. The classic

work of Box and Cox (1964) offered a way, for a large class of problems, to let data indicate an appropriate transformation, but they were at pains to stress that advice from the data is not always to be taken literally. An estimated power of 0.1 signals the use of logarithms to researchers of any experience!

Harry Crane (*Rutgers University, Piscataway*)
I applaud Gelman and Hennig's advocacy for subjectivity in statistical practice and appreciate the overall attitude of their proposal. But I worry that the proposed virtues will ultimately serve as a shield to deflect criticism, much like objectivity and subjectivity often do now. In other words, will not acceptance of 'virtue' as a research standard in short order be supplanted by the 'pursuit to merely *appear*' virtuous?

I believe Gelman and Hennig when they assert '[W]e repeatedly encounter publications in top scientific journals that fall foul of these virtues' (Section 6.2). I am less convinced, however, that this 'indicates [. . .] that the underlying principles are subtle'. This conclusion seems to conflate *doing* science and *publishing* science. In fact I suspect that most scientists are more or less aware of these virtues, and many would agree that these virtues are indeed virtuous for doing science. But I would expect those same scientists to acknowledge that some of these virtues may be regarded as vices in the publishing game. Just think about the lengths to which journals go to maintain the *appearance* of objectivity. They achieve this primarily through peer review, which promises transparency (virtue V1), consensus (V2) and impartiality (V3) but rarely deliver any of them. It should be no surprise that a system so obsessed with appearances also tends to reward research that 'looks the part'. As 'communication is central to science' (Section 2.3) and publication is the primary means of scientific communication, is it any wonder that perverse editorial behaviours heavily influence which virtues are practised and which are merely preached?

Finally, I ask: just as statistical practice is plagued by the 'pursuit to merely *appear* objective', is science not also plagued by the pursuit to 'appear statistical'? Judging from well-publicized issues, such as *p*-hacking (Gelman and Loken, 2014; Nuzzo, 2014; Wasserstein and Lazar, 2016), and my own conversations with scientists, I would say so. To borrow from Feyerabend (2010), page 7, 'The only principle that does not inhibit progress is: anything goes'. So why should we not simply encourage scientists to make convincing cogent arguments for their hypotheses however they see fit, without having to check off a list of 'virtues' or to run a battery of statistical tests?

Wasserman (2012) invited us to imagine 'a world without referees'. Instead, I am envisioning a world without editors, journals or statistics lording over science and society: without 'objectivity' obscuring the objective, and without 'virtues' standing in the way of ideals. That world looks good to me.

David Draper (*University of California, Santa Cruz*)
I would like to reinforce the point made in this admirable paper about the crucial role of problem context, and to suggest a notational method that may help to keep context at the forefront of our thinking, where it belongs. Problems $\mathbb{P} = (\mathbb{Q}, \mathbb{C})$ in which statistical methods are helpful are defined by the real world questions $\mathbb{Q}$ that they pose and the context $\mathbb{C}$ in which those questions are raised. The pair $(\mathbb{Q}, \mathbb{C})$ in turn identifies a triple $(\boldsymbol{\theta}, \mathbf{D}, \mathcal{B})$ of important ingredients: $\boldsymbol{\theta}$ is the unknown of principal interest, $\mathbf{D}$ is the available data set that You (Good (1950); a person wishing to reason sensibly in the presence of uncertainty) can use to decrease your uncertainty about $\boldsymbol{\theta}$, and $\mathcal{B} = (B_1, \ldots, B_k)$ is a finite set of (true or false) propositions, all rendered true by problem context, that exhaustively summarizes the relevant background information about how $\mathbf{D}$ was gathered, together with all other assumptions and judgements arising from $\mathbb{C}$ that aid in the telling of a reasonable probabilistic story about how $\boldsymbol{\theta}$ and $\mathbf{D}$ are related. For instance, if a randomized controlled trial has been performed, yielding $\mathbf{D}$, then one of the B_i might be '(patients were randomized with equal probability to treatrnent or control and followed for n weeks)'. In the Bayesian paradigm, in which quantities on either side of the conditioning bar can be either sets or propositions, it then becomes natural to condition explicitly all probability statements on $\mathcal{B}$ (and frequentists could use notation such as $P_{\mathcal{B}}(Y > c)$. As a Bayesian example, $p(\boldsymbol{\theta}|\mathcal{B})$ quantifies all ('prior') information about $\boldsymbol{\theta}$ external to $\mathbf{D}$ that is implied by $\mathbb{C}$; context is sometimes sufficient to identify a unique prior (e.g. Laplace's *principle of indifference* is actually a theorem when the only prior information about $\boldsymbol{\theta}$ is that it can take on the m values $(v_1, \ldots, v_m)$) but, if You are making modelling choices that are not driven by problem context in prior specification, this can be made explicit in your notation (e.g. $p\{\theta|\mathcal{B}, U(0,1)\}$ for a uniform prior on $0 < \theta < 1$). See Draper (2013) for examples of this notational approach in real world problem solving. I conjecture that the accurate and explicit use of $\mathcal{B}$—including encouraging people to say precisely what propositions are contained in their $\mathcal{B}$s—would yield greater clarity in statistical modelling and communication.

David Firth (*University of Warwick, Coventry*)
I congratulate Gelman and Hennig for providing such an insightful view of some principles of our discipline.

The paper proposes seven desirable attributes that are difficult to argue with. A statistical analysis ideally should be transparent, be based on consensus, be impartial, correspond to observable reality, be aware of multiple perspectives, be context dependent and be stable.

To these I would suggest adding one more: 'A statistical analysis should ideally be *done*'.

What I mean by this is that, too often, statistical principles are used as an excuse for *not* drawing inferences, or—worse still—as a reason for a statistician to say that another's method of analysis is invalid but without actually providing a better answer.

The discipline of statistics rightly emphasizes such attributes as relevance and correctness of conclusions drawn from an analysis. At its heart, though, statistics is an *enabling* discipline. With this in mind, our applied work will most often be appreciated by other scientists when we *can* produce an answer, or when we *can* help to improve an analysis done elsewhere. Principles such as those discussed in the paper will surely help to guide us; but ultimately, in the words of the authors, we often just have to 'do our best'. Doing a useful analysis, and *in time* for it to be genuinely useful, should be an overriding aim.

Simon French (*University of Warwick, Coventry*)

Firstly, I thank Gelman and Hennig for not just an enjoyable and stimulating paper, but a paper that addresses the foundations of what we do. The paper echoes discussions from the 1960s, 1970s and 1980s, a time when we lacked the computational power to analyse data properly, so we discussed by what criteria we would judge an analysis proper. Nowadays, it seems to me, too many analysts compute without much thought to the process of inference that they are putatively supporting. Clearly Gelman and Hennig do not do that. I support the approach that they promote wholeheartedly. I say that as a card-carrying subjective Bayesian who recognizes that subjectivism gives *me* a perspective to understand my analyses, but to work with others I need the more operational concepts: the *virtues* that the authors propose.

However, I think that their arguments would gain if they considered a larger context than statistical analysis to support scientific inference. My allegiance to the Bayesian model stems from the coherence that it provides in supporting inference *and* decision, and my professional life has focused on risk and decision analysis. Those areas bring to the fore two issues that the authors skim past.

Firstly, Gelman and Hennig's examples focus on *statistical* models with hierarchies, clusters and the like, though almost as an aside they do mention challenging, computationally expensive models in pharmacology. In many risk and decision analyses, the statistical models are negligible in complexity compared with the *physical* models which predict the consequences. Those, often controversial, physical models carry much of the prior knowledge into the analysis, not just through their parameters, but through their structure. There is much more here than a simple choice of a regression model (point (f) in Section 1). How do the arguments extend to circumstances when the structure of a physical model provides much of the prior knowledge?

Secondly, their discussion broadly assumes that they are data rich relative to the inference being made. In many risk and decision analyses, directly relevant data are sparse and one needs to draw in experts to provide judgemental data. Although Cooke (1991) has made some suggestions, we have yet to develop sound means of reporting expert-judgement-based studies (French, 2012). What suggestions do Gelman and Hennig have?

Adrian Gepp (*Bond University, Gold Coast*)

I commend Gelman and Hennig on their proposal to shift the debate away from objective *versus* subjective methods and instead to focus on the virtues of both approaches. I would like to pick up on the authors' comment that, in the machine learning domain (Section 5), there is 'more emphasis on correspondence to observable reality (virtue V4)'. Friedman (1997) demonstrated that more accurate probability estimates do not guarantee improved classification, and in fact often lead to worse classification. This counterintuitive result was eloquently explained by Hand (2009) many years later. As an example, a fraud detection model may perform better at classifying cases as fraudulent or legitimate by using biased, rather than unbiased, estimates for the probabilities that fraud had occurred. Consequently, classification tasks in machine learning may be improved by using probability estimates that have less, rather than more, correspondence to the observed reality (virtue V4). However, in some cases the classification probabilities might be more important than the classifications themselves, such as when ranking cases in terms of their likelihood of fraud. Hence, the best approach and the relative importance of virtue V4 depend on the purpose of the machine learning task, which demonstrates the importance of virtues V5 and V6 about context dependence and multiple perspectives and how they can relate to other virtues.

Jan Hannig (*University of North Carolina at Chapel Hill*)

Gelman and Hennig should be congratulated on their bravery in writing a philosophical contribution. Although I do not agree with all the points I believe a discussion of what makes a 'good' statistical analysis is very desirable.

I wholeheartedly agree that transparency is very important. It is closely related to another virtue—reproducibility; the ability for others to reproduce the results in the paper on their own from the information and data provided. However, in my mind objectivity goes well beyond that to what some call replicability; the situation when another team performing a similar experiment, using the same analysis, would arrive at a similar conclusion. Replicability is related to the last of the virtues listed by the authors: stability; but I feel that it goes far beyond that to ideas such as avoiding overfitting.

I was intrigued by the virtue of considering multiple perspectives. There is a need to develop tools for statisticians to present how different conclusions and their uncertainties may be depending on the strength of our assumptions. The uncertainty pyramid of Lund and Iyer (2017) seems to be a first step in this direction.

Frequentist are often viewed as more objective with less choices to make whereas Bayesians are viewed as making wore. The authors are correct to push against this viewpoint. I would like to add that the nature of the Bayesian paradigm of modelling everything with a single probability distribution leads to the conclusion that once a suitable model and prior have been selected all possible questions are uniquely answered. This is fine when very strong informative priors are used but, as the literature on reference priors points out (Berger *et al.*, 2009), weakly informative priors must be tailored to questions of interest. From this point of view frequentist methods can be viewed as more flexible because, by the nature of the frequentist paradigm, procedures need to be tailored to the parameters of interest.

Finally, Liu and Meng (2016) have talked about the robustness *versus* the relevance trade-off. In their language robustness is the control over the properties of the procedure in an idealized setting—a frequentist concept. Relevance is the idea that conclusions should be valid for data sets resembling the one at hand—a feat that Bayesian conditioning achieves perfectly. One of their hypotheses is that fiducial inference often strikes a good balance between these two competing goals. As someone who spent some effort developing generalized fiducial inference (Hannig *et al.*, 2016), I fully agree that fiducial approaches are definitely worth considering in the quest for honest statistical inference.

Giles Harper-Donnelly (*University of Cambridge*) **and Peter Donnelly** (*University of Oxford*)

How refreshing it is to see a paper on foundational–philosophical issues in statistics which is rooted in practicality and does not come from the perspective of knowing all the answers. Instead it acknowledges the very real challenges of doing serious statistical analyses in complicated settings. Being open and honest about this must be a good thing.

The discussion of transparency brought to mind a quote attributed to Jack Good that

> 'The subjectivist (i.e. Bayesian) states his judgements, whereas the objectivist sweeps them under the carpet by calling assumptions knowledge, and basks in the glorious objectivity of science',

which had an important influence on the statistical journey of one of us (PD). In fact Good argued for a constant interplay, rather than a polarization, of subjective and objective perspectives (Good, 1983). Happily the tenor of modern discourse is less dogmatic and pejorative (and less gendered!) than it was 40 years ago. What then seemed an unbridgeable ideological chasm between frequentists and Bayesians is routinely traversed in modern statistical applications, as this paper so aptly illustrates.

In the areas of science in which we work—developmental biology (GHD) and genetics (PD)—statistical analyses often require methods developed specifically for the application, rather than any standard toolkit. A major practical problem is that scientists will try methods when helpful easy-to-use software is available, without necessarily understanding, let alone checking, critical assumptions underpinning the method, or *caveats* needed in interpreting results. Routine documentation of these (explicit *and* implicit) assumptions would be very helpful.

There is a concerted push by many scientific journals towards greater transparency in research, including an insistence on giving details of sample sizes and statistical procedures used. This is mainly driven by concerns about reproducibility and is of course to be welcomed. This paper would suggest, and we would strongly agree, that in addition we should push for moves to facilitate 'assessability': including specific statements and justification of key assumptions, including those that are implicit in choices of procedures or criteria, alongside an explicit critique of the models used.

We heartily commend the authors on their paper, and on their broader and very effective efforts to

encourage good statistical practice. If we take seriously the sophisticated application of statistics in complex scientific settings, then this paper should become a landmark in the field. The openness that it promotes should also become a standard part of the way statistics is taught.

Jack Jewson (*University of Warwick, Coventry*)

I thoroughly enjoyed how this paper brings to light the subjectivity disguised as objectivity in statistical practice, and I relish the prospect that understanding the impossibility of objectivity will allow researchers greater freedom.

Focusing on the Bayesian standpoint, Gelman and Hennig's discussion omits examining methods for parameter updating. They regard Bayesian updating to be objective and transparent, suggesting that if their prior is interpretable then so is their posterior inference. In the M-closed world I can believe that this is so. However, in the M-open world (Bernardo and Smith, 2001), where the authors acknowledge that the majority of statistics takes place, Bayesian updating is less transparent. Bayesian parameter estimation is known to learn about the parameters of the model that minimizes the Kullback–Leibler divergence to the data-generating process, but in practical terms I do not believe that many statisticians understand what it means to be close in terms of Kullback–Leibler divergence. The general Bayesian update (Bissiri *et al.*, 2016) reinterprets Bayes's rule as follows:

$$\pi(\theta|\mathbf{x}) \propto \exp\left\{-\sum_{i=1}^{n} -l(\theta, x_i)\right\}\pi(\theta) = \exp\left[-\sum_{i=1}^{n} -\log\{f(x_i;\theta)\}\pi(\theta) = \pi(\theta)\prod_{i=1}^{n} f(x_i;\theta)\right]. \tag{1}$$

This demonstrates that greater posterior mass is given to parameter values whose predictions via the model $f(\cdot;\theta)$ achieve a low logarithmic score on the observed data $\mathbf{x}$, providing greater transparency. Bernardo and Smith (2001) observed that scoring predictions based on the logarithmic scoring rule places great importance on correctly specifying the tails of the data-generating process, but in applied problems the statistician may require their predictions to be accurate in some other region. The authors acknowledge that 'how the results of an analysis are to be used or interpreted' provides important subjective information. If the tails of the predictive distribution are important, then the logarithmic score should be chosen and this decision should be documented. However, if the tail specification is not important, then (implicitly) minimizing the logarithmic score can produce predictive distributions that perform very poorly on the rest of the distribution.

In this scenario it is tempting to implement the general Bayesian update without using a model. I agree with the sentiments of the authors that using a model is important; it provides another tool to incorporate prior information in the analysis and induces transparency in the way predictions are produced. Research is therefore being conducted into model-based loss functions allowing Bayesian updating to target aspects of the predictive distribution away from the tails.

In agreement with the authors' recommendations concerning priors and tuning parameters, I advocate that Bayesian updating cease to be considered an objective black box and the room to impose subjectivity is exploited and documented.

Julie Josse (*École Polytechnique, Paris*)

Although I agree with Gelman and Hennig that virtuous statistical practice involves justifying choices made during the analysis, I do not think that statisticians do not do it because it is subjective, but rather because no one cares sufficiently. Even if such explanations are crucial, they are not valued by the community. It is not common to have an entire paper on the topic of scaling (see Bro and Smilde (2003)), and such papers are likely to be published in applied journals not perceived to be prestigious by other colleagues. The pressure to publish should be mentioned.

I do not think either that there are endless discussions on the subject of objectivity and subjectivity, but the fact that there are many ways to deal with a problem will always lead to this impression of subjectivity. This debate seems more linked to the Bayesian literature, perhaps because it has at least the merit of questioning what information is incorporated in the analysis. This could explain why those who use it for mathematical simplicity, which is quite justifiable, may be seen as 'opportunistic Bayesians'. It is crucial to make choices clear.

The choice of data coding is important. In sensory analysis, there is debate about whether the Likert scale should be coded as quantitative or qualitative. To be coding free, some methods (Pages, 2015) consider a compromise between these two points of view and highlight the specificity of each. The example of clustering is striking. Callahan *et al.* (2016) also stressed the need to document analyses with a view to reproducibility. He has shown that there could be 'more than 200 million possible ways of analyzing these

data'. Of course, there is no 'good' solution; it depends on the characteristics of the data one wants to capture.

Even when a problem is well characterized, two statisticians who make use of the same data will use different approaches. This is mainly due to their personal history (on a personal basis, I speak about 'a statistical sensitivity'), and the expression 'when one has a hammer, one sees nails everywhere' often applies. This is not necessarily a problem, and experience gained must be used. As mentioned by the authors, collaboration should be encouraged, e.g. in the development of simulation studies.

In conclusion, this paper has the merit of promoting transparency, awareness of the limits of a study and its context-dependent nature. (As a French statistician, I also appreciated that the paper begins with 'we cannot do statistics without data.) The battle is not lost because the community is already encouraging the sharing of code and data. It is worth remembering that different points of view can be legitimate. 'The best future is one of variety not uniformity' (John Chambers).

Saana Jukola (*Bielefeld University*)

According to Gelman and Hennig, the words 'objective' and 'subjective' are not used in a helpful way in statistics. They argue that, instead of references to objectivity and subjectivity, scientific practices should be evaluated by using a broader collection of attributes.

I agree with them that, instead of subjective–objective duality, it is more fruitful to use a more refined criterion for evaluating scientific endeavours. This is because, first, as the concepts of objectivity and subjectivity have several meanings (Douglas, 2004), calling something objective can lead to misunderstandings. Second, given the rhetorical force of 'objectivity', it can be epistemically—or even ethically—detrimental to name a process or its outcome 'objective' or 'subjective' without reflecting what assumptions and ideals concerning the knowledge production are embedded in this evaluation (e.g. Jukola (2017)). Consequently, being more precise about which features of inquiry are seen as beneficial to its goals can help to justify and assess scientific practices.

As the authors mention (Section 6.2), their own list of virtues that we could use instead of 'objective' or 'subjective' (Section 3) for evaluating different methods and projects in statistics may not be fully applicable or sufficient to be used in some other scientific fields. In particular, it might be that in some contexts the list should be supplemented with an attribute or attributes that address the institutional or social conditions of the inquiry, e.g. the presence of potential conflicts of interest regarding the source of funding. The works by social epistemologists (e.g. Longino (1990)) who have previously developed accounts for evaluating how the broader social system influences knowledge production and the conditions for acquiring reliable information could be taken into account here.

For a philosopher of science, this paper is a valuable contribution. On the one hand, it has potential to deepen previous philosophical analyses of 'objectivity' (e.g. Douglas (2004) and Hacking (2015)) by illustrating how assumptions that often are ingrained in the use of the concept and its counterpart 'subjectivity' can influence practical scientific work. On the other hand, it shows how philosophical theorizing can be relevant for actual scientific practice.

Kuldeep Kumar (*Bond University, Gold Coast*)

First I congratulate Gelman and Hennig for a thought-provoking paper and for igniting the debate on subjective and objective issues in statistics. As Silver (2012) has pointed out in his famous book the debate between subjective and objective issues is the debate between 'Bayesian' and 'frequentist' approaches in statistics. Historically Ronald Fisher was worried with the notion of 'Bayesian priors' as it seemed too subjective and cuts against the notion of objective science. In this paper the authors have addressed this controversial issue by listing various virtues of a good scientific or statistical investigation. I think that one of the important virtues from a regression or forecasting point of view is virtue V4(b) 'clear conditions for reproduction, testing and falsification'. In this context, in my view there should be a balance between objective and subjective methods, and I would prefer to create a third hybrid category 'semisubjective' or 'semiobjective'. Quite often we must combine the two approaches especially when making an out-of-sample forecast or forecasting for a new product (for which no prior data are available). A forecast can be generated by using quantitative methods as a means of objectification but then subjective judgement must be applied to obtain the final forecast. Silver (2012) pointed out that decision makers look at probabilistic (subjective) information and translate that into a decision (objective) which is deterministic. He also mentioned that we can never achieve perfect objectivity, rationality or accuracy in our beliefs, but we can always strive to be less subjective, less irrational and less wrong (Silver (2012), page 259).

The other area that is not considered in the paper is sampling techniques, which play a key role in statistics. Subjectivity in statistical problems arises because quite often we deal with samples rather than

the whole population. A representative sample of the population can be drawn more efficiently by using a combination of objective sampling techniques (random sampling) as well subjective sampling techniques (judgemental sampling).

Manuele Leonelli (*University of Glasgow*) **and Jim Q. Smith** (*University of Warwick, Coventry*)

We congratulate Gelman and Hennig on a thought-provoking paper that puts forward a set of virtues which any statistical analysis should enjoy. We believe that these have the potential of guiding practitioners in critically questioning many of the arbitrary choices currently made in applied analyses, often obscured under justifications of 'objectivity'.

Of course, however, there is a big difference between a practitioner making assessments for her own purposes, e.g. how to invest her money on the stock market, and how she can on behalf of a company present information about the safety of a product to convince a regulator (bringing what are *legitimate* data and reasoning as evidence). Although in the first case above a practitioner may not need or even want to obey the virtues proposed, documentation reporting all the steps and assumptions of a statistical analysis addressing these virtues becomes critical in the latter. However, the paper does not seem to appraise this fundamental difference.

We have observed the necessity of analyses enjoying the virtues of transparency and consensus in complex heterogeneous systems consisting of chains of models, each under the jurisdiction of different panels of experts, where the output of one model is used as input for following models. The outputs of these separate models then need to be pasted together coherently to provide an overall assessment of the whole complex system and to rank the policies examined by a decision centre. However, in this context it is not enough to demand only coherence in its formal sense (Bernardo and Smith, 2000), but it is also essential to be able to provide a narrative which is sufficiently plausible to encourage a third party scrutinizing the model to accept its probability statements.

Integrating decision support systems (Leonelli and Smith, 2015; Smith *et al.*, 2015) are a framework that carefully selects the outputs of each component model and then combines these appropriately via tailored propagation algorithms. For these systems a supporting narrative explaining the policies scores can be composed around the rationales delivered by the different panels, thus addressing the virtue of transparency in massive complex systems. This overall distributed narrative then provides a platform around which policy makers can discuss the evidence supporting one policy against another. On the basis of this platform, assessments of policies can be discussed and, if needed, revised: an interactive capability recognized vital to any such integrating decision support system to build consensus.

Nick Longford (*Imperial College London*)

Gelman and Hennig should be commended on this venture into a topic that is central to how we (should) think and operate in research and practice of statistics as a scientific discipline and profession. Although prominent throughout the paper, I think that objectivity and subjectivity are not discussed in their full complexity, as they relate specifically to statistics. Without doubt, all science strives to be objective and should have the virtues listed in Table 1. However, our agenda—what we study and how we set our priorities and distribute our efforts—is subjective, controlled by our interests, capacities and external incentives.

I suggest that statistics has some commonalities with the medical and legal professions, namely a commitment to serve the best interests of the client in making decisions in the presence of uncertainty. This aspect is unjustly relegated in references to statistics as a science that aims to establish associations and mechanisms from imperfect (incomplete) information or data.

I have expressed my dissatisfaction with how we manage the uncertainty in this effort, by estimation (with minimum mean-squared error), hypothesis testing (with a set test size), quoting confidence intervals (with a prescribed rate of coverage) and other common formats of inference (Longford, 2013, 2016), with model selection mixed in (Longford, 2012). These formats are oblivious to the perspectives, value judgements and priorities of the client, who would be well advised to weigh the consequences of the errors that are committed by such inferential statements. I am referring here to errors that arise not as a result of professional incompetence or misconduct.

The client will receive a better service from us if we integrate these perspectives in the analysis. An easy target for criticism is Section 4.4, where the analysis is conducted for a distant (hypothetical) client whom we do not care to understand or construct a plausible perspective. A careful study of the client's perspective may be rewarded by a solution that is more closely tailored to the interests of the client. Such a solution is unequivocally subjective. But the methods for eliciting the perspective are (or should be) objective. An objective perspective is a compromise of (many) subjective perspectives, and so the corresponding solution

Fig. 2. Simple schematic diagram to illustrate the dependence of the range of plausible values of a quantity of interest and increasingly strict requirements for a set of assumptions to be considered plausible: the degree of restrictiveness increases from left to right along the horizontal axis; at any given point on the horizontal axis corresponding to a particular set of assumptions, the vertical line segment bounded by the broken curves shows the range of plausible values for the quantity of interest

serves these perspectives well only on average. Most perspectives would be served better by permitting the 'heresy' of different, even conflicting, solutions that are specific to the various (plausible) perspectives.

Steven Lund and Hari Iyer (*National Institute for Standards in Technology, Gaithersburg*)

We applaud Gelman and Hennig for their clear and constructive articulation of the need to de-emphasize subjective *versus* objective arguments and, instead, to strive for transparency, stability and the other virtues they enumerate when considering alternative proposals for modelling and analysing any given set of data.

Modelling choices are often defended or explained as satisfying some chosen criteria. In most cases, alternative models will also satisfy these criteria, and by exploring the range of results attainable by those models one can begin to understand, in an organized manner, the effect of information inserted via assumptions. This general framework can be pursued in Bayesian and frequentist analyses alike.

The *relationship among data, assumptions and interpretation* (RADAI) is generally nebulous. Rather than seeking to provide a single range of values (e.g. the posterior credible interval or confidence interval) for the characteristic of interest, seeking to illuminate the RADAI may be more in line with the virtues listed. Fig. 2 provides a simple schematic comparison between conventional uncertainty characterizations (the vertical line) and the target of RADAI analysis (the broken curves indicate uncertainty ranges across increasingly strict plausibility constraints). For simplicity, consider the analysis of a given data set. To illuminate the RADAI, a statistician might define

(a) a characteristic of interest, X, as a function of an unknown distribution,
(b) a space of plausible distributions to consider at the outset, and a corresponding range of attainable values for X, and
(c) a plausibility criterion evaluating the compatibility between candidate distributions and their assumed relationship to available data (used to prune the initial set of distributions considered, after which the range of attainable values for X may correspondingly decrease).

Examining the range of interpretations available under different initial sets of considered models (representing assumptions regarding the unknown distribution), and different plausibility (e.g. goodness-of-fit) criteria dictating how observed data will be used to filter an initial set, may constitute the most complete and transparent portrayal of the RADAI a statistician can offer. We refer to this framework as a 'lattice of assumptions' leading to an 'uncertainty pyramid' (Lund and Iyer, 2017).

O. J. Maclaren (*University of Auckland*)

Summary

Andrew Gelman and Christian Hennig propose that the concepts of objective and subjective, in the context of statistical practice and statistical foundations, are better thought of as *complementary* concepts rather than as strictly opposing concepts. They further suggest that they are typically better replaced by more specific and interesting terms which nevertheless preserve some useful kernel of the original distinction.

Response

I agree! I similarly almost always find the terms objective and subjective used in unhelpful ways, and better replaced by alternative pairs of concepts rather than completely discarded. Their table of 'virtues' has the potential to facilitate much more productive discussions of statistical foundations.

Andrew and Christian also mention some threads lying beyond the usual Bayesian–frequentist divide that they hope discussants will pick up on. Here are two I find interesting.

Exploratory data analysis: models and procedures

Exploratory data analysis and robustness concerns represent developments in statistics with important implications for statistical foundations. How these implications are interpreted depends in part, however, on whether we adopt a model-based or procedure-based perspective. As Huber and Ronchetti (2009) state (page 325):

> 'The differences between... model-based and... procedure-based approaches surfaced in [an]... interchange between... George Box and... John Tukey.... In Tukey's view, robustness was an attribute of the procedure, typically to be achieved by weighting or trimming the observations. Box, on the other side, contended that the data should not be tampered with... the model itself should be robust.'

I would be interested in the authors' thoughts on this apparent tension, especially as, I believe, Christian is closer to the Tukian spirit and Andrew the Boxian spirit. For example, when should we attempt to *model* 'noise' and when should we *process* 'noise'?

Uncertainty: statistics and parameters

Statistics usually distinguishes between observable quantities (data and associated statistics) and unobservable quantities (parameters and models themselves). This distinction is, to me, strongly related to that, mentioned briefly by Andrew and Christian, between 'known unknowns' and 'unknown unknowns'.

To bring all sources of uncertainty 'within' the purview of probability theory, Bayesians must blur the distinction between statistics and parameters. The price is requiring additive measures over all quantities, including 'unobservables'. This implies, formally, that only one parameter set (or model) of those under consideration can be 'correct'. What do the authors think of this trade-off?: in particular, implications for identifiability issues and/or the need to consider non-probabilistic or non-additive uncertainties.

Jean-Michel Marin (*Université de Montpellier*), **Julie Josse** (*École Polytechnique, Paris*), **Christian P. Robert** (*Université Paris Dauphine, University of Warwick, Coventry, and Centre de Recherche en Economie et Statistique, Paris*)

The focus in the paper is definitely set on statistical models and the ensuing inference. We wonder whether this focus is set on an inappropriate problem, namely arguing about the best way to solve the *wrong* problems, whereas most users are more than ready to settle for *ad hoc* solutions, provided that these carry a minimal *modicum* of efficiency. In short, there are many more immediate production problems that shout for statistical processing than well-set scientific questions like the existence of an elementary particle.

We adhere to the argument that the scientific realism position enables a more workable *modus operandi* This particularly applies to data analyses in medical and social sciences, as opposed to hard sciences where (almost) all experimental conditions can be expected to stay under control or at least to be stationary across repeated experiments. Maybe not so incidentally, the three examples treated in Section 4 belong to the former category. (We find it quite significant that Gelman and Hennig picked a clustering example as this certainly is one statistical procedure loaded with preconceptions and forking interpretations.) These examples are all worth considering as they feed, albeit in specific contexts, the authors' arguments. However, they give the impression that the major issue does not truly stand with the statistical model itself, referring instead to a concept that is relevant only for hard sciences. This was further illustrated by the mention of outliers during the talk: a non-sensical notion in an M-open perspective. It is obviously illusory to imagine that the *all models are wrong* debate settled but it would have been relevant to see it directly addressed. Therefore, the dismissal of machine learning in Section 5 is disappointing, because it is worth considering at least for loosening the reliance on a background model, using mostly predictive performances. The alluring universality of such tools, plus the appearance of objectivity produced by the learning analogy, could have been addressed, especially in a context when these techniques are overtaking more traditional learning in many areas.

Finally, the effect of software should be mentioned. The implementation of some methodologies may explain why certain practices, possibly flawed, are still favoured. Although software does require default values for tuning parameters, users should be made aware of the underlying choices and validate them. Furthermore, all-inclusive statistical solutions, used by innumerate practitioners possibly inappropriately, give them the impression of conducting 'the' statistical analysis. This false feeling and its relevance for this debate also transpire through the mishandling of statistical expertises by media and courts, and some scientific journals.

Jorge Mateu (*University Jaume I, Castellón*)

Gelman and Hennig are to be congratulated on this valuable contribution and thought-provoking paper dealing with a timely and extremely interesting topic involved with the statistical practice and statistical thinking in basically all areas of science, but also lining up with a more philosophical attitude to the concepts of objectivity and subjectivity in science. My point for discussion is far more philosophical than perhaps mathematical, trying to express the link between these two concepts: the nature of the statistical analysis and the open area of data science we are currently facing. I start by restating the authors' feeling that these encountered attitudes about objectivity and subjectivity can be an obstacle to good practice in data analysis and its communication.

As I see things, when we deal with mathematical arguments to develop theoretical concepts in statistical science, we can be much more objective, separating ourselves from subjective interpretation of the results. However, even in this more clear case, when assumptions are imposed on the theory to be true, we are somehow growing closer to subjective interpretations, and pragmatic decisions must be taken even in this more theoretical context. But, data are in the roots of statistical science and in the corresponding statistical analysis, and a general accepted point is that ideal objectivity can never be achieved. Thus, I cannot understand statistical analysis and personal decision making without the intervention of the researcher with a corresponding dose of subjectiveness. I strongly disagree with those who consider statistics to be the science of defaults.

Data come as the new source of information power world wide. Their real potential is the business insights that can be derived from this new, vast and growing natural resource. If data are the next big thing, then researchers and companies need to think about a new business model that exploits this valuable resource. To extract the most value from data, we need to be able to analyse, process and integrate an extremely large amount of new 'big data' types in almost realtime. We thus need new technologies and processing capabilities, and to develop new analytic skills to gain more insights and to evolve from a traditional reactive to a new predictive approach. The latter objective is actually the real challenge and defines a new motivation for the new data science discipline. Grounded and sound science can only be developed by amounts of the right combination of subjectiveness and objectiveness.

Marina Meilă (*University of Washington, Seattle*)

I would like to point out a relationship between *stability* and *correspondence to observable reality.* The latter brings to mind *goodness of fit*. I shall briefly show how stability together with goodness of fit can validate the inferences we make without reference to truth, but only to observations. If observed data are 'what is not going away when we [change our beliefs]', and truth is not to be talked about, what is a model, then? It is a formula that encodes the inferences we make from the data.

For example, when we employ a linear classifier $\hat{y}_\theta(x) = \mathrm{sgn}(\theta^{\mathrm{T}} x)$, with $x \in \mathcal{X}$, $\theta \in \Theta$ and $\mathcal{X}, \theta \in \mathbf{R}^p$, the model $\hat{y}_\theta$ encodes all the labels that we shall assign by this procedure. When we obtain a K-clustering $\mathcal{C}$ of data by Lloyd's K-means algorithm (Lloyd, 1982), the quadratic loss is our implicit model of how groups in data should be.

Stability can validate such inferences. For instance, in the case of the linear classifier above, learning theory (Vapnik, 1998) bounds the probability of error by the empirical error probability $(1/n)\Sigma_{i=1}^{n} \mathbf{1}_{y_i \neq \hat{y}(x_i)}$ plus a term that depends on n and the Vapnik–Chervonenkis dimension of the function class $\{\hat{y}_\theta, \theta \in \Theta\}$. Let this bound be ϵ. Hence, with high probability over the sample, all predictors $\hat{y}_\theta$ with low empirical error will infer about the same labels, with variations of at most 2ϵ between them. This *stability result* and the bound 2ϵ enable us to trust the inferences, without reference to truth.

Stability bounds exist in the clustering example as well. Given a clustering $\mathcal{C}$, Meilă (2006) showed that, under conditions that depend only on the data and $\mathcal{C}$, any clustering $\mathcal{C}'$ with $\mathrm{loss}(\mathcal{C}') \leqslant \mathrm{loss}(\mathcal{C})$ cannot differ by more than ϵ from $\mathcal{C}$, where the difference is measured by a distance between partitions and ϵ depends on $\mathrm{loss}(\mathcal{C})$.

In both cases above, the *stability bound* depends on the goodness of fit of the model and is informative only if this fit is good. Goodness-of-fit measures and distribution-free results guarantees in statistics are not always studied from this perspective, or even together. The aim of this note is to encourage this point of view.

Angela Montanari (*University of Bologna*)

I found the paper by Gelman and Hennig full of stimulating and original ideas. While reading it, I tried to figure out how the virtues that they propose and describe can find room in the topics that I usually study and work on. In particular, among others, I focused my attention on fitting a linear factor model

for exploratory purposes. Most of the fitting steps can be done according to virtues V1, V4 and V7, for instance in the decision of whether to work on raw data or on standardized data, in the choice of the estimation method both for the factor loadings and the factor scores, in the determination of the number of latent factors. But something a little puzzling happens when decisions need to be taken regarding factor rotation. It is well known in fact that the linear factor model is rotation invariant; this implies that there is an infinite number of equivalent factor models that might generate the same data and therefore that there is no data-driven criterion that can enable us to choose the 'best' one. The issue is indeed common to any not identifiable model as the linear factor model is.

Bartholomew *et al*. (2008) wrote

> 'A criticism often levelled against factor analysts is that subjectivity seems to play a big role. This is a misreading of the situation in two respects. First, one cannot obtain any solution one wants. Secondly, the situation is more accurately described as one in which the same solution can be expressed in different ways.'

This quote looks like a nice rewriting of virtue V5, i.e. awareness of multiple perspectives. I would be interested in Gelman and Hennig's opinion on this point. Do they believe that in this case V5 is still a virtue? And, if they do, how do they suggest we exploit this virtue in a fully exploratory context? Furthermore, do they also deem unidentifiability a virtue anyway? I am slightly concerned that we are making a virtue of necessity.

Matthew Moores (*University of Warwick, Coventry*)

Statistics is an essential element of scientific practice and, as such, statistical procedures should be evaluated with regard to the philosophy of science. Towards this goal, Gelman and Hennig propose seven statistical virtues that could serve as a guide for authors (and reviewers) of scientific papers. The chief of these is transparency: thorough documentation of the choices, assumptions and limitations of the analysis. These choices need to be justified within the context of the scientific study. Given the 'no-free-lunch' theorems (Wolpert, 1996), such contextual dependence is a necessary property of any useful method.

The authors argue that 'subjective' and 'objective' are ambiguous terms that harm statistical discourse. No methodology has an exclusive claim to objectivity, since even null hypothesis significance testing involves the choice of the sampling distribution, as well as the infamous $\alpha = 0.05$. The use of default priors, as in objective Bayes methods, requires ignoring any available information about the parameters of interest. This can conflict with other goals, such as identifiability and regularization. The seven virtues are intended to be universal and can apply irrespectively of whether the chosen methodology is frequentist or Bayesian. Indeed, the authors advocate a methodology that combines features of both.

The main feature of the methodology in Section 5.5 is iterative refinement of the model (including priors and tuning parameters) to fit the observed data better. Rather than Bayesian updating or model choice, the procedure suggested involves graphical summaries of model fit (Gelman *et al*. (2013), chapter 6). It has connections with calibrated Bayes (Dawid, 1982) and hypotheticodeductive Bayes (Gelman and Shalizi, 2013) methods. This is an excellent approach, albeit saddled with an unfortunate misnomer.

The term 'falsificationist' seems misleading, since this would imply abandoning an inadequate model and starting again from scratch. As stated by Gelman (2007),

> 'The purpose of model checking (as we see it) is not to reject a model but rather to understand the ways in which it does not fit the data'.

Leaving aside the question of whether statistical hypotheses are falsifiable at all, except in the limit of infinite data, falsification in the Popperian sense is really not the goal. Furthermore, this approach is not limited to posterior predictive distributions. It could be applied to any generative model, whether such a model has been fitted by simulation (e.g. Monte Carlo sampling) or optimization (e.g. maximum likelihood or cross-validation).

Fionn Murtagh (*University of Huddersfield*)

The themes of this paper are very important for the contemporary context that focuses so much on data science and on 'big data' analytics.

In addition to what is noted in Section 2.1 about John Tukey, it is good to note how relevant also is the work of Jean-Paul Benzécri, who emphasized that 'The model should follow the data, and not the reverse!'. Interestingly the same title as Tukey (1962) also holds for Benzécri (1983).

So much in this paper is of importance in our current context, for teaching and learning, for research, for decision making and for inductive reasoning. For many aspects of data science and big data analytics, see for example Murtagh (2017).

Very key themes which are having major current research carried out include, in Section 3.2, multiple perspectives and context dependence, sparsity for computability and interpretability (see Murtagh (2016)), the very major importance of Section 4.1 for the inherent hierarchical nature of very high dimensional spaces and of all that is complex in nature (see Murtagh (2017)). In Section 4.2, the role of transformation is very relevant for the resolution scale of our analytics. There is much to be gained from the homology and field concepts in the work of the eminent social scientist Pierre Bourdieu (see such discussion in Murtagh (2017)).

Much of Section 5 is well aligned with the correspondence analysis platform for data analysis. Through the incorporation and indeed integration of exploratory data analysis, extended to all that is covered by unsupervised data mining and knowledge discovery, with, wherever relevant the adjunction of statistical modelling, supervised classification and all that is associated with these domains, this leads to our analytics being declared to be inductive inference. Thus there is the informal expression, that we are 'letting the data speak'.

The final paragraph of Section 4.3 states the need for us, in our statistically based analytics, not to be 'ignoring context and potentially multiple perspectives'. This is good and relevant for taking one's stand, in the often-used statement, that 'correlation is not causation'.

Keith O'Rourke (*O'Rourke Consulting, Ottawa*)

Whereas few statisticians have done more than dip their toes into the philosophy of science, Gelman and Hennig have strolled well in and brought back some edifying insights and exhortations for the statistical community. Of course, the statistical community is not that interested in philosophy of science *per se.* Nor, would I suggest, should it be. Rather it should want a comprehensive grasp of how best to make empirical inquiry profitable and to keep it profitable: profitable, that is, in the sense used by C. S. Peirce (both a philosopher and applied statistician among other things) who argued that aesthetics (what to value or set as a goal) informs ethics (how to act and interact) and ethics informs logic (how best to inquire).

Now, how to act and interact to make empirical inquiry profitable, Peirce would argue, should follow only from the desire to find out how things really are—the goal to be valued above all, i.e. getting at that 'observer-independent reality' as Gelman and Hennig put it, which has no direct access but which we truly hope our being able to act without being frustrated is actually connected with. Gelman and Hennig's virtues (ethics) perhaps should be evaluated and weighed (informed) with respect only to the desire to find out how things really are. For instance, consensus (virtue V2) is not an end in itself but rather just a profitable means to find out how things really are. Reasonable attempts (or perhaps better bending way over backwards) to achieve consensus may lead to unexpected realizations of how to become less wrong. In isolation or weighted too heavily (as Gelman and Hennig do point out) it is a serious vice rather than a virtue. The same holds for *investigation* of stability (V7), though stability as a goal I would argue is misguided. Given more space, I believe each virtue could be likewise assessed to clarify how they should inform statistical practice, i.e. with appropriate weighting they will and with inappropriate weighting they will not.

My own efforts to obtain a comprehensive grasp of how best to make and keep empirical inquiry profitable have largely been based on reading C. S. Peirce and regrettably few other philosophers. This is primarily a historical accident and continued convenience, as I find his work continually helpful. With that, I would point out Gelman and Hennig's referenced source for active realism in turn referenced C. S. Peirce as the main source of those ideas (Peirce Edition Project, 1998).

L. A. Paul (*University of North Carolina at Chapel Hill and University of St Andrews*)

Gelman and Hennig replace the goal of discovering an objective, mind-independent, scientific account of reality with the goal of forming a consensus based on multiple, diverse, rigorously developed scientific points of view. The idea seems to be that, at the limit of such consensus, we can reach an understanding that is independent of particular points of view. In this way, we achieve a kind of observer independence.

This view of science as a consensus-based collection of multiple, diverse, points of view requires us to specify how to move in a stable and rigorous way from the perspective of the individual scientist to a representation of reality that quantifies over a range of perspectives. Good science does not involve the elimination of so-called 'subjective' decisions and assumptions. Rather, it requires a transparent and rigorous assessment of them.

I agree. The problem is deeper than commonly appreciated. The need to understand and clarify the role of subjective choices and assumptions about data collection and experience has implications for decision making (Paul, 2014) and inference from evidence, up to and including evidence from the 'gold standard' of randomized controlled trials (Paul and Healy, 2017).

As we assess different potential contributors to our consensus-based representation, it is essential to attend to the types of choices and judgements that relate to theory construction and data analysis. For example, we need to recognize and attend to how the development of a scientific point of view encodes aims and assumptions, expressions of confidence and approximation, and individual inventiveness.

I depart from Gelman and Hennig, however, when they suggest that truth is no longer an aim. Mathematical and logical truths form the foundation for good science. Context dependence is not opposed to truth. (A standard approach is to hold that the truth of a claim is indexed to its context. For more on contextual dependence and truth see, for example, DeRose (2009) and Cappelen and Dever (2016).) Further, to distinguish between what counts as rigorous (or non-rigorous) requires us to have a basis for that distinction. Unless we are willing to replace discovery of the truth with mere instrumental success, we need a notion of rigour as one that cleaves closer to the true than to the false. The value of the consensus-based, rigorously vetted, scientific representation of the world is that it gives us a view of reality that is closer to the truth than its competitors.

Emilio Porcu (*University Federico Santa Maria, Valparaiso, and University of Newcastle*), **Bernardo Lagos Alvarez** (*University of Concepción*) **and Libia Lara Carrión** (*Universidad Andres Bello, Viña del Mar*)

Objective or not, statistics was born to be used.

We congratulate Gelman and Hennig for their beautiful paper about emigration of different areas of knowledge. Such an emigration should not cause any conflict between disciplines. The paper brings opportune conceptualizations, the virtues, that enable us, until now only approximately, to conjugate a homologation of future developments in statistics and its related areas, to the updating of a contextualized language and with a greater understanding of its scope.

Thus, the proposal contributes to classifying the relevant problems of the superfluous and their respective methods to attack such problems. This is of paramount importance in the communications era, along with advances in the devices developed for this communication to be massive, of data or information, whose patterns to the naked eye of the independent observer of reality, are diffuse.

We must take care that decision making is not the product of complex strategies of communication placed in practices by interests which are not devoted to the progress of humanity. We believe that, to put into practice the previous proposal, we should not leave aside the development of the scientist or professional in the attributes that must be analysed, as described in the proposal. In addition, we risk being naive, in favour of the consensus, that an attribute of 'statistical institution' should start to be cultivated.

Jan-Willem Romeijn (*University of Groningen*)

The paper by Gelman and Hennig contains a stimulating discussion on the intersection of statistics and the philosophy of science. Considering the pivotal role of statistics in all empirical sciences, and the prominence of the scientific method as a topic in the philosophy of science, these two disciplines have much to offer to each other. In what follows I hope to illustrate that, by providing a further context for the virtues that Gelman and Hennig discuss.

It is customary in the philosophy of science to distinguish between aleatory and epistemic conceptions of probability. The ideal of objectivity can be split in a similar fashion, referring either to the virtue of correspondence to facts and eventually truth, or to the virtues of transparency, impartiality and, more generally, rationality. This distinction maps roughly onto the distinction between ontology and epistemology, between what there is and what we know. And, as I argue below, it clarifies the respective roles of the virtues that replace objectivity.

Focusing on epistemology, both impartiality and transparency are aspects of a virtue that to my mind merits independent representation: *logical validity.* The idea is that statistical inference answers to an independently motivated and indeed fully objective rationality norm, namely probabilistic coherence. This core virtue ensures impartiality because it purges the inferences of substantive assumptions, and transparency because it brings those assumptions out as explicit premises. The virtue of logical validity thus forces statisticians to acknowledge their modelling assumptions, and the perspective taking that it involves. Falling back on the traditional vocabulary for effect: it is precisely by making the subjective starting points of statistical inference explicit that we can truly maintain its procedural objectivity.

Moving from the rationality of statistics to the truth or falsity of its substantive results, note that logical validity is defined as the conservation of truth through inference. Depending on the specifics of the decision situation, researchers will take different assumptions as input to their statistical inferences. What matters to objectivity in the metaphysical sense, i.e. in the sense of delivering true results, is whether those assumptions are true. Accordingly, nothing in the dependence of statistics on modelling assumptions, or in the variety of perspectives adopted by researchers, prevents us from arriving at true conclusions about the matter at hand. Viewing statistics as a logic helps us to clarify what it takes to arrive there, and also saves us from relativism about what statistics can deliver.

Jan Sprenger (*Tilburg University*)

I congratulate Professor Gelman and Dr Hennig for a paper that bridges philosophical and statistical inquiry in an impressive way. They deserve credit for taking subjective judgement out of the closet, and for showing how it improves statistical reasoning when applied in a transparent, open-minded and empirically informed fashion. This connects well to philosophical analysis of the social dimension of objectivity: transparency about subjective assumptions facilitates scientific discussion, opens up avenues for mutual criticism and promotes the formation of a balanced, informed judgement (Longino, 1990; Harding, 1991; Douglas, 2009; Sprenger, 2017a).

The paper is also notable for what it does *not* say: that objectivity corresponds to value freedom, and that the influence of data on theory can be assessed without reference to values. A long research tradition in the philosophy of science has found this ideal unattainable, especially in the context of inductive reasoning (e.g. Rudner (1953), Hempel (1965), Douglas (2000) and Reiss and Sprenger (2014)). With regard to the role of values in inference, I invite the authors to expand on the virtue of impartiality (Table 1, virtue V3). Does it involve priority of evidence over values, or balancing values against each other, as Douglas (2004, 2009) has suggested?

The rest of this note is devoted to removing a common misunderstanding about Bayesian inference. Gelman and Hennig write that Bayesian inference should not be limited to the analysis of subjective beliefs (Section 6.1). I agree, but I would like to be more radical and to deny that Bayesian inference should be understood like this in the first place. Probably no Bayesian statistician thinks that the prior distribution over different parameter values mirrors her *actual* degrees of belief (see also Gelman and Shalizi (2013)). Rather, the prior formalizes the degrees of belief she would have on the *supposition* that the overarching statistical model is true. Of course, all these models are highly idealized and most probably false. The prior and posterior distribution should not be understood as determining betting odds on the truth of the various hypotheses; rather, they express *relative plausibility judgements* conditional on a given model (Sprenger, 2017b).

Like anywhere in science, these judgements are only as reliable as the model from which they are derived: 'garbage in; garbage out'. The 'falsificationist Bayesianism' discussed in Section 5.5 flows naturally from taking the subjective Bayesian approach seriously. It makes explicit that inferences within a model need to be complemented by checking the adequacy of the model itself. And I agree with Gelman and Hennig that this critical perspective is vital for Bayesians in pursuit of scientific objectivity.

Milan Stehlík (*Johannes Kepler University in Linz and University of Valparaíso*)

Congratulations go to Gelman and Hennig for giving readers a provocative discussion on subjectivity and objectivity.

I would like to point out that there is no real need to replace 'objectivity and subjectivity' with broader collections of attributes. Actually, neither objectivity is replaceable by 'transparency, consensus or impartiality' and 'correspondence to observable reality', nor should 'awareness of multiple perspectives' and 'context dependence' replace subjectivity. Of course, in special instances, the newly introduced attributes could give some information in a relevant set-up, but replacement is not justified. It will be worth investigating closer relationships between introduced attributes and, in the current state of the art, subjectivity and objectivity are very useful concepts. From the formal view you replace 'two levels' by 'multilevel structure'. To put it in the order of mathematical logic, you are trying to build 'modules in statistics classification', which is the subject of many-valuedeness in formal logic. For the concept of clustering much inspiration can be found in the Poincaré paradox (see Höhle (1990)).

Before we start to do any statistics, we shall justify its fitness with respect to information theory. Thus subjectivity and objectivity of statistics are important topics. See for example Stehlík *et al.* (2017), where it is illustrated that for example the choice of Kolmogorov's axiomatic theory of probability was developed as one possible alternative, but in many experiments, e.g. in physics and finance, we should go beyond these axiomatics.

Aside discussion on axiomatics and logic, underlying phenomena can have chaotic behaviour, which cannot be captured by means of statistics (see Stehlík *et al.* (2016, 2017) in the context of climate change). In such data contexts outcomes of statistics are truly subjective.

Finally I have one question for poets: have you ever met a statistician who is transparent, impartial, has a sense for consensus, likes correspondence to observable reality, has awareness of multiple perspectives and masters context dependence? I have met several subjective ones already.

Stephen M. Stigler (*University of Chicago*)

No one in recent years has done more than Andrew Gelman and his colleagues to point to the difficulty in coping with the effect of unacknowledged steps taken in an investigation, in choosing what leads to pursue, what data to include and what questions to be asked. He calls this the 'garden of forking paths', borrowing a title from a Borges story, and he has forcefully made us aware of how that can destroy the validity of a naive significance test, leading to at least some of the noted lack of reproducibility of scientific research. It is not an altogether new idea: in 1885 Alfred Marshall wrote

> 'The most reckless and treacherous of all theorists is he who professes to let facts and figures speak for themselves, who keeps in the background the part he has played, perhaps unconsciously, in selecting and grouping them'

(Stigler (2016) page 202). But the problem is easily forgotten and requires constantly renewed attention.

Gelman and Hennig in effect argue that a change in terminology will help. I am not convinced. They have valid criticisms of 'objective' and 'subjective', but at least we have long experience with them, and it is not clear that the replacements suggested do not suffer from similar problems: they have not been subjected to the same searching examination by generations of philosophers and statisticians. They anoint the new terms as 'virtues', but have these earned that moral status, or are they each, as Hacking said of 'objective', merely types of lack of vices? All the new terms carry their own baggage in the form of past usage and misusage. 'Impartiality' is usually loaded with unannounced bias; major gains in science have come from challenging the 'consensus'; 'transparency' is rapidly becoming a meaningless buzzword. And how do we judge the degree of 'correspondence to observable reality'?: by the P-value of a χ^2-test? A salesman may find 'context dependence' and 'multiple perspectives' better than 'subjectivity' in closing a deal, but are they more clearly understandable?

The authors are correct to believe that 'objective' and 'subjective' are problematic terms in statistics. Their arguments are cogent; their examples clear; their point well taken. The authors' aims are meritorious and their analysis of the problems insightful, but their solution, to my view, merely moves the problem down the road, to an unpaved section with unfamiliar potholes. Statistics is difficult, and a simple name change will not alter that.

Mervyn Stone (*Ruislip*)

Gelman and Hennig have been able to condense their many nuances into a bold table of impeccable virtues. As a checklist, Table 1 could play a valuable role in discussions about statistical arguments, but I would ask whether it can add much to their resolution when specific arguments are vigorously debated in their social context. It may be that its rather onerous commandments should be kept out of sight of potential Fellows tentatively thinking of a statistical career.

My unease in some frankly political contexts is not so much with the *quality* of reasoned argument as with its *absence*. Will the authors please explain how their checklist could have helped in achieving consensus about arguments when simply responding to their perceived flaws failed to do so? In the case of a recently used National Health Service funding formula, what could a checklist have usefully added to fully documented exposure of the formula's reliance on spurious tests of model specification (Galbraith and Stone, 2011) or to exposure of the ignorance of regression to the mean on the part of stout defenders of the formula (Williams and Stone, 2013)? The same question can be posed for the recently proposed police funding formula given that exposing the glaring irrationality of fixing resource allocations by the weights of the standardized cost variables in the first principal component has failed to elicit any response from the Home Office. There are other examples of flawed arguments for Table 1 to be tested on in the Word Press bundle WhitehallStatisticalReasoning.wordpress.com.

Mauricio Suárez (*Complutense University of Madrid*)

The debate over the nature of probability is about a century old and shows no sign of abating: do probabilities represent objective features of the world, or the incomplete knowledge of cognitive agents? If the

former, there are objective chances; if the latter, there are subjective credences or states of partial belief. Traditionally, these two answers have been thought to be incompatible. Probability may represent chance, or it may represent credence; but it cannot represent both.

Gelman and Hennig argue that statistical modelling involves both objective and subjective elements—and, rather than concentrate on what is objective and what subjective in our statistical models, we would be better served by considering the functional virtues that these models have.

I agree wholeheartedly. In my own work (Suárez, 2017) I have been emphasizing the role that judgement plays in statistical modelling. Far from following any algorithmic or automatic procedure, statisticians routinely make finely balanced decisions to establish the appropriate sample space that characterizes a particular phenomenon. That they must exercise judgements does not imply that the practice is merely subjective or dependent on an arbitrary whim, any more than say aesthetic appreciation, physical intuition or moral judgement are not whims. On the contrary, statistical judgement can be very normatively constrained. One improves at it with practice. Gelman and Hennig are right to point out that judgement does not entail subjectivism. On the contrary, relative to any particular context, there are always objectively better or worse judgements.

Whether this abolishes the distinction between objective chance and subjective credence is to my mind doubtful—and I have defended the role for objective chance in statistical practice—but it certainly provides us with a much more faithful description of the ways in which both interlink in statistical practice.

Peter F. Thall (*University of Texas MD Anderson Cancer Center, Houston*)

Gelman and Hennig provide a stimulating discussion of practical and epistemological problems with use of the words 'subjective' and 'objective' in statistics. As an alternative, they propose that statisticians pursue seven virtues, which they describe in detail.

One might make an analogous list of 'seven deadly sins' frequently committed, by both authors and editors, in the published medical literature. This issue seems related to virtues V4 and V7, which encompass model criticism. A very common sin is reporting a data analysis based on a model assumed purely for convenience or by convention without any assessment of how well it fits the data at hand, or consideration of alternative models. It might be very useful to promulgate the authors' recommendations to the communities of non-statisticians whose knowledge of statistics often is limited to how to run statistical software packages.

Considering consensus to be a virtue implicitly assumes that the opinions of all involved individuals are sensible. In clinical trials, pursuing consensus often reinforces undesirable conventional behaviour. Examples in oncology research include the common use of single-arm clinical trials as a basis for treatment comparison, cherry-picking data to support preconceived beliefs and intentionally choosing not to publish results of clinical trials or laboratory studies that do not provide 'statistically significant' results.

Attention to the seven virtues should aid in design of simulation studies to obtain frequentist properties of complex Bayesian or frequentist methodologies, and possibly to calibrate tuning parameters on that basis. Unfortunately, in much of the statistical literature, details regarding simulation study design, choice of statistics used to evaluate performance and numerical methods often are missing.

The remark in Section 5.2 that frequentist inferences typically assume a fixed model while ignoring prior model selection brings to mind the fact that reported p-values or confidence intervals from a fitted frequentist regression model seldom are correct. The frequentist distributions of parameter estimates from such fitted models depend on the model selection process. Deriving corrected numerical values by bootstrapping or cross-validation typically reveals a painful truth, but this seldom is done. Given that p-values are used to quantify strength of evidence, this common frequentist convention is sorely lacking in virtue.

With their seven virtues, the authors have set the bar very high. Perhaps virtue in statistics is a goal to be pursued, if not always perfectly achieved. I thank the authors for their thought-provoking paper.

Tyler VanderWeele (*Harvard T. H. Chan School of Public Health, Boston*)

Gelman and Hennig are to be commended for articulating criteria that more precisely capture notions of objectivity and subjectivity in the use of statistics. Certainly the virtues of transparency, consensus, impartiality, correspondence to observable reality and stability are all essential and ought to be discussed with great frequency. As much as I appreciated their paper, I do have three critiques to offer.

First, their virtue of consensus seemed to capture two distinct ideas: first that it is possible to attain consensus on evidence and conclusions and, second, that researchers follow communally established common procedures. The first form of consensus relates closely to what might be taken as knowledge within a research community. The second form of consensus concerns method (Lonergan, 1965); consensus about

common method is helpful in the progress of and attaining knowledge consensus within science, even if common method must sometimes be challenged and is occasionally overturned (Kuhn, 1962). However, common method is only one source of consensus on knowledge within science. The various other virtues—transparency, impartiality, conformity to observable reality and, perhaps especially, stability—all contribute critically. I would thus suggest as the virtue 'common method' and suggest further that 'consensus' (in knowledge) is not so much a virtue within statistics as a central marker or indicator of objectivity to which all the virtues contribute.

Second, concerning the supposed virtues of 'awareness of multiple perspectives' and 'awareness of context dependence', such *awareness*, although desirable, is too weak. The question is how the multiple perspectives are *handled.* Are they investigated? And, critically, are the results stable across multiple methods (or when changing priors)? Sensitivity analysis is imperative here (Oakley and O'Hagan, 2004; Greenland, 2005; Lash *et al.*, 2009; Ding and VanderWeele, 2016). It is when conclusions are consistent across methods or perspectives that we attain consensus. This is only sometimes, not always, possible. However, once we move to questions about how multiple perspectives and context dependence are *handled*, we are arguably simply back to the virtues of transparency and stability.

Finally, I would dispute the paper's characterization of science in the statement, 'finding out the truth about objective reality is not the ultimate aim of science ... science rather aims at supporting human action'. Prominent scientists throughout history have seen things very differently. If only use is of concern, many of the proposed 'virtues' could effectively be ignored. While striving in some sense to clarify the notion of objectivity in statistics, it seems the paper has slipped into a relatively extreme position on the subjectivity of truth. This does science no favours. Our capacity to arrive at truth may be limited, but to give up the quest entirely is arguably an abandonment of science itself.

Wolf Vanpaemel and Francis Tuerlinckx (*University of Leuven*) **and Paul De Boeck** (*Ohio State University, Cleveland, and University of Leuven*)

We agree with Gelman and Hennig that debating the subjectivity or objectivity of a study or method is fruitless and wasteful, and we welcome their introduction of a more helpful set of criteria, which constitutes a useful checklist for doing good science reporting. We would, however, like to extend the focus of their proposal to include also the design of studies. Personal decisions and arbitrary choices not only abound in data processing and analysis, but also during data collection (and the choices then even become crystallized in direct replication studies).

The arbitrary choices involved in data collection seem especially important in our field, psychology, in which the object of study, and thus the size of the effect, is inherently variable and context dependent (i.e. sensitive to specific circumstances). The heterogeneity in effect size is real, as one can observe in Fig. 1 in Klein *et al.* (2014), which shows a lower bound to the real heterogeneity (because maximal uniformity of the studies was aimed at).

Even in the simplest data collection efforts, there are many design features to be decided. For example, a researcher must make the following decisions when using ratings to capture the dependent variable: what are the exact instructions? In which order will the rating questions be presented? In which format, on paper or on line, with or without time limit will the questions be presented? How, if at all, will participants be rewarded? What is the exact content of the cover story? Often, there are many different reasonable implementations, and singling out just one is a seemingly arbitrary decision.

The currently dominant approach involves reporting an $N = 1$ sample from the larger population of possible meaningful studies. As an alternative, we propose to conduct a metastudy, which is a collection of many small studies. Each study is constructed by randomly sampling from the various design options, and participants are randomly assigned to a study. A metastudy forces researchers to be explicit about the range of application of an effect and sheds light on the generalizability of the effect. Depending on the size of heterogeneity, metastudies have more power and are thus more efficient compared with the traditional meta-analysis of large uniform replication studies. Further, metastudies are a way to deal with a reality that may be too complex for large theory-based moderator studies. Consistent with some of Gelman and Hennig's criteria, metastudies explicitly acknowledge the awareness of multiple perspectives in data collection, and directly investigate the stability of a finding.

Damjan Vukcevic, Margarita Moreno-Betancur and John B. Carlin (*Murdoch Children's Research Institute and University of Melbourne*)

We thank Gelman and Hennig for their fresh take on fundamental principles of statistical practice. Although we agree with the proposed virtues, we suggest that more emphasis should be given to a strengthened

version of virtue V6, which we suggest naming 'clarity of purpose'. A premise of applied frequentists and Bayesians alike tends to be that the primary task of the statistician is to build a model, often parametric, that approximates the 'true' data-generating process (Section 5.5). But we would ask where do such models come from and are they relevant or meaningful absent a prespecified purpose?

The authors recognize that models should be built with a focus on the aim of the analysis but may not emphasize sufficiently that models are tools rather than ends in themselves. Focusing on obtaining the 'right model' may distract from or even hinder achieving the underlying aim. An example is if the ultimate goal is to estimate a *single* causal effect or parameter. The emerging discipline of causal inference (e.g. Pearl (2009)), particularly the strand of 'targeted learning' (van der Laan and Rose, 2011), suggests that here it may be worth putting more effort into clearly specifying a meaningful target parameter, along with (non-parametric, external) causal identifying assumptions, and then estimating the target by using methods that avoid oversimplified and unnecessary parametric assumptions about other aspects of the data.

Furthermore, when thinking beyond philosophical debates to the real world of statistical practice, it is important to ask: whose practice? Most statistical analyses are actually performed by non-statisticians (Section 2.1), whose practice is dominated by procedural approaches ('which test should I use?') with 'objectivity' as an implicit ideal. These analyses are typically used to provide a 'stamp of approval' for claims of scientific discovery or 'truth' (Section 2.2). Such mechanical usage is inconsistent with most of the virtues, which appropriately emphasize a more nuanced view of statistical analysis as a means of evaluating and quantifying the uncertainty surrounding hypotheses under investigation, rather than for making black-and-white claims (Gigerenzer and Marewski, 2015; Goodman, 2016). So although we think that this paper is a great step forwards for statisticians—who definitely need to move beyond 'finding the best model' and 'the most powerful test'—it will be important also to engage more with the practice that is done in the name of statistics but not by statisticians.

Eric-Jan Wagenmakers (*University of Amsterdam*)

A visit to Disneyland can produce mixed feelings. Yes, most rides are great, the fireworks are spectacular and the kids love it. But then there are also long queues, poor restaurants and the general annoyance caused by other people. The paper evokes sentiments that are similarly mixed. The great rides deserve first mention. The plea to abandon the terms 'objective' and 'subjective' is well taken. In fact, in 1963 the archetypical objectivist Harold Jeffreys discussed the work of archetypical subjectivist Jimmy Savage and stated 'I have suggested that "subjective" has been used with so many meanings that it should be discarded as hopelessly confusing' (Jeffreys (1963), page 408).

In addition, the emphasis on model uncertainty is apt, as is the emphasis on the dangers of data contingent analysis choices (e.g. Peirce (1878)). I applaud the call for adding external information to the statistical model, and I fully agree that the specification of the likelihood deserves at least as much scrunity as the specification of the prior distribution. Moreover, I appreciate the suggestion that different analysis methods may be legitimate and nonetheless yield different answers. These are key insights, which are often overlooked, that empirical disciplines would do well to incorporate in their statistical curricula. Presently, students still come away from statistics courses falsely believing that there is a single correct analysis that yields a single all-or-none conclusion. Humans abhor uncertainty, but statisticians and scientists should steel themselves and accept the inherent ambiguity that arises whenever a finite set of data is analysed.

The paper also contains some long queues and poor restaurants. I fail to grasp the flirt with frequentism: a non-evidential paradigm that prides itself on having slaughtered common sense on the altar of objectivity. If one wants to learn from prediction errors and not to go empistemically bankrupt, there is a simple alternative: stick to the tenets of Bayesian probability theory.

Moreover, I was mystified by the description of the subjective Bayesian. It is true that a subjective Bayesian cannot switch to a different model, but only if this subjective Bayesian first violates Cromwell's rule and deems all other models impossible to begin with. Finally, the authors argue that a prior distribution cannot be evaluated. I beg to differ. Statistical models are a combination of likelihood and prior that together yield predictions for observed data, the adequacy of which can be rigorously assessed (as is commonly done by using Bayes factors).

Priyantha Wijayatunga (*Umeå University*)

Subjectivity can rarely be avoided in statistics since almost all analyses are done by people who have some subjectivity. However, analysts should have some general objectiveness and must be capable of explaining

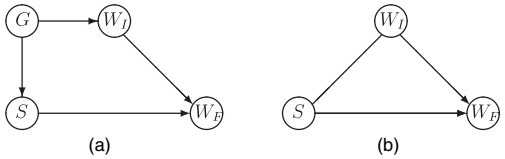

Fig. 3. Causal diagrams for Lord's weight gain problem: model (a) is the underlying model of model (b); they can be used to calculate the causal effect of S on W_F (or equivalently on D) with standard theory detailed in Pearl (2009)

and convincing their solutions to anyone in the problem domain, answer their (often subjective) questions and show how theirs are consistent across different subjective opinions, etc.

Inherently regression models are subjective and cannot be objective unless their prediction accuracies are of concern. Take so-called Lord's paradox (Lord, 1967); when weight gain D is modelled with gender S by $D = \alpha_0 + \alpha_1 S + \epsilon$ then α_1 is not statistically significant whereas, when initial weight W_I is included in the model, $D = \beta_0 + \beta_1 S + \beta_2 W_I + \epsilon'$, then β_1 is significant. Two subjective opinions are encoded as two linear regression models that are correct as long as their residuals ϵ and ϵ' follow the required assumption (they do so), i.e. as long as their predictions are of concern. Regression coefficients generally represent predictive effects, not causal effects that are obtained under stricter conditions. A paradox arises when interpreting them as causal effects.

In contrast with regression models, causal models should be objective since they assess real effects. However, since they are also done by people, there can be some subjectivity in them as well. The solution given in Pearl (2016) for the paradox is one instance, where initial weight W_I is taken as an intermediate variable on one causal pathway from S to D (equivalently final weight W_F). Alternatively, one can argue that the weight is associated with the gender but not causally affected by it, i.e. the person's genetic information G determines the person's gender and the weight, among other personal characteristics. They are just effects or features of the genetic information. In the absence of the genetic information, for modelling weight at a later time, one can take both S and W_I as causal factors of it, i.e. all the genetic information for determining the weight at a later time is represented by the initial weight and the sex, i.e. $G \perp W_F | (W_I, S)$ (Fig. 3). This argument yields a different result from that of Pearl (2016); two subjective opinions result in two different answers to the same causal question.

Robert L. Winkler (*Duke University, Durham*)

This paper proposes some attributes that are more helpful to think about than subjectivity and objectivity when conducting statistical analyses and making decisions. These attributes are important and deserve careful consideration.

The subjective–objective dichotomy in statistics has its roots in the Bayesian–frequentist debate that seemed most heated when Bayesian methods were starting to gain traction in the 1950s–1970s. To my mind, it is a false dichotomy. As Gelman and Hennig suggest, all aspects of statistical analysis, from framing the question of interest to building statistical models to collecting, analysing and interpreting data, are ultimately subjective. Choices are being made by the statistician all along the way. Even different followers of 'objective Bayes' methods attacking the same problem can generate quite different analyses via different data-generating models or different 'objective' priors. The availability of competing classes of 'objective priors' belies any notion of objectivity. My objection here is not to the methods, but to the term 'objective'.

Some feel very strongly about the dichotomy, but is it still of major concern? For a quick non-scientific check, I searched the papers in a recent issue of the *Journal of the American Statistical Association* (volume 111, issue 516). Of 31 articles, only three used 'subjective' or 'objective', just mentioning 'subjective probability estimates' or 'objective priors' in a matter-of-fact way. For the most part, Bayesians and frequentists appear to coexist peacefully these days.

The attributes suggested in the paper are appealing. To the experienced statistician, they provide a nice road map of important aspects to consider. They are commonsense notions that are like 'motherhood and apple pie', to use an American expression. That is, they should be considered appealing by most people and are hard to criticize. Sadly, things like transparency, consensus, impartiality and even correspondence to observable reality seem to be disappearing in an increasingly polarized society. We can see that in our

combative social discourse and in the proliferation of 'fake news' or 'alternative truth' that can change from day to day.

In practice, many statistical analyses are conducted not by highly sophisticated, experienced statisticians, but by analysts with limited statistical training using standard software. Their concern is finding a convenient method to apply to their problem, not about philosophical issues, and they may be under time pressure. Will these ideas and the extra time they require impact practice only for sophisticated statisticians working on large-scale problems, or can they become part of standard statistical practice at all levels? What are the implications for the teaching of statistics?

Michael J. Zyphur (*University of Melbourne*) **and Dean C. Pierides** (*University of Manchester*)
We agree with Gelman and Hennig that researchers should avoid naively conceived notions of 'objectivity' and 'subjectivity' (e.g. Good (1989), chapter 4). We also appreciate replacing these terms with 'virtues'. Yet, we have three concerns. First, objectivity and subjectivity are opposed in ways that *also* trouble their surrogate virtues. 'Observable external reality' is equivalent to 'consensus' in terms of what a community believes is observable (e.g. dark matter or not and social aggregates as real *versus* constructed), and therefore these virtues may be opposed to 'multiple perspectives' and 'context dependence'. The authors' approach works as if science is confined to a community in agreement. Yet, for revolutions or major discoveries, and in all the places wherein dissensus exists, the authors' virtues may be contradictory, ignoring what is practically possible for researchers. The authors' virtues are communal *outcomes* that enable normal science and therefore may not help to understand the contested spaces that make science interesting.

Second, in Appendix A the authors gloss over a rich body of scholarship that addresses our first concern and offers a robust understanding of how communities require virtues and values such as 'objectivity' and 'subjectivity' to organize their research, crucially by shaping both scientific practices and their practitioners (for brief work, see Barad (l996, 1998), Hicks and Stapleton (2016), Porter (1991, 1992, 1993) and Shapin (2016); for thorough treatments, see Barad (2007), Shapin (1994, 2008), Porter (1995) and Poovey (1998)). For example, conceptions of what constitutes reality and how to study it are coproduced with ways of institutionally forming and orienting scientists. Key elements are how to navigate areas wherein agreement does not exist about the boundaries of things like subjectivity and objectivity.

Third, our primary concern embodies a classical pragmatist focus on outcomes and ethics: how should we evaluate reformulations of scientific concepts and practices? It is not enough to be virtuous within a discipline, just as it is not enough to forward notions of virtues without engaging with the wider communities that science should serve (Dewey, 1938). The authors note that their work is 'ultimately justified on pragmatic grounds', but pragmatic in relation to whose means and ends, which worldly problems and their solutions, and which ethic(s)? If reality is a social construction as the authors propose, then wider communities and their problems may be a good place to find evaluative concepts to replace 'objectivity' and 'subjectivity' when assessing the value of statistical and other scientific practices.

The **authors** replied later, in writing, as follows.

As practising scientists it is natural to feel that we are too busy for philosophy. We are grateful to have found so many thoughtful discussants who agree with us on the value of reflecting on the values, often left unstated, which underlie statistical theory, methods and practice.

We wrote our paper because we believe that words matter. From one direction, we feel that outmoded attitudes of objectivity and subjectivity have impeded the work of statisticians, by discouraging some good practices and by motivating some bad practices. At the same time, we recognize the real concerns underlying the decades-long debates regarding objective and subjective perspectives in statistics, and we believe that an uncovering of these more fundamental concerns, as represented by the list of virtues in Table 1 of our paper, can help us to do better by freeing ourselves from artificial restrictions and by suggesting new directions.

But there are precedents for the expansion of the philosophical frameworks of statistics. Consider exploratory data analysis, which was traditionally considered to be separate from academic statistical methods but which has been brought into the fold via the formalizations of Tukey (1977), Rubin (1984), Gelman (2003), Wilkinson (2005) and Wickham (2017). It is not just that we now have a language to talk about statistical graphics in the context of models; we also have many valuable tools which allow us to learn from, to check and to improve our models in ways that were not accessible when graphics were taken merely as good practice without a theoretical structure. For a completely different example, the framework of missing data has been used to systematize causal inference (Rubin, 1974). More recently, the replication

crisis in psychology has stimulated new thinking regarding the interplay between statistical inference and experimental design in the context of the scientific publication process (Simmons *et al*., 2011; Button *et al*., 2012).

The common theme in all the above examples is the parallel development of critiques of existing practice, ideas for improvements and philosophical or theoretical developments. We view this discussion as a step in this process: our paper was an attempt to jolt the ideas of statistical objectivity and subjectivity into the modern world, and the discussions are a necessary course correction. Indeed, this conversation exemplifies several of the virtues that we tabulated in our paper, including full communication (virtue V1(c)), openness to criticism and exchange (virtue V3(c)) and awareness of multiple perspectives and context dependence (virtues V5 and V6). The largest challenge in such a discussion may be correspondence to observable reality (virtue V4). We spelled out connections to statistical practice in Section 4, and we hope to keep these and other applied examples in our minds in this discussion and moving forward.

We shall go through the 53 (!) discussions in what seems to us to be a logical order, recognizing that brevity stops us from being able to respond to all the issues that arose in the discussion.

We are delighted that the discussion contains many valuable additional suggestions. The list of meta-statistical considerations provided by Dawid is particularly rich and deserves a paper on its own.

We begin with reactions to our general set-up. Vanpaemel, Tuerlinckx and De Boeck point out that statistical discussions are typically framed in the context of a single study or a single analysis, but that statistics by its nature is concerned with ensembles. In any example, we should consider our design and analysis choices in the context of other problems that we might be studying. This is a frequentist idea to consider the properties of statistical methods relatively to a reference set, and it is also Bayesian in that the elements of this set can be assigned probabilities, to form a prior distribution or hierarchical model representing a larger class of problems.

McConway and Cox, Thall, and Porcu, Alvarez and Carrión, connect Dawid's metastatistics to another meta-idea: that the principles of transparent, aware science are also relevant to scientific communication and publication. We agree with Longford that the reporting of inferential summaries such as point estimates, uncertainties and hypothesis tests involves choices which should be made with awareness of applied goals and with understanding of who will read these summaries and how they should be used.

Also related is the point made by Grant and Firth, and by Harper-Donnelly and Donnelly, that transparency is great, but transparency plus explanation is even better. Any good statistical analysis comes with scientific explanation—and, again, the virtues that are discussed in our paper (which in turn are derived from introspection and reflection on what we view to be good statistical practice) are relevant to the explanation of quantitative findings. Future progress on exploratory model analysis (Urbanek, 2006; Wickham, 2006) may benefit from a more careful elaboration of goals and trade-offs among our list of virtues, for example following up on Morey's remark that increasing selectivity may decrease sensitivity. From the other direction, Josse, alone and with colleagues, argues that a lack of clarity in explanation could be a natural response of researchers to the incentives of publications that favour the illusion of certainty and objectivity. As Suarez writes, one gets better at statistical judgement with practice, and it helps to work within a framework that allows for such experimentation. One motivation for writing our paper was to engage the philosophy community in this interplay we see between the philosophy and practice of statistics.

Stigler and Stehlík argue that the terms 'subjectivity' and 'objectivity' actually do convey opposing approaches to learning from data and thus we have no need to discard these in favour of our longer list of virtues. From a purely intellectual perspective, we might agree; but, as noted above, we feel that arguments surrounding these terms have polluted the discourse in statistics, and so we prefer to focus on what we consider to be more fundamental goals. Our list of virtues in many settings is aspirational rather than descriptive; then, again, 'objectivity' is aspirational also, and we believe that our aspirations are more understandable, more realistically attainable and more useful in grappling with the hard problems of statistics. We do, however, agree with Stone that 'honesty and transparency are not enough' (Gelman, 2017) and we did not mean to imply that complete adherence to our list of virtues—even if this were logically possible—would resolve all or even most statistical problems. There will always remain many challenges in computing, mathematical understanding of statistical methods, communication, measurement and many other statistical problems. What we hope to gain from our framework is a clearer view to help us to avoid various roadblocks that have arisen from slavish following of outdated philosophical principles.

For brevity, some issues received less attention from us than they deserve. Major examples are non-parametric statistics, machine learning and the M-open perspective in Bayesian statistics (as highlighted by Robert, Marin and his colleagues and Meilă). Our intention was certainly not to dismiss these approaches, which are attractive in many applications, particularly where prediction is a major aim.

Model assumptions are often presented in statistics in a misleading way, as if the application of methods would require assumptions to be true. But reality will typically differ from formal model assumptions (see also Hennig (2010)). Models set up idealized situations in which methods can be shown to work well. In this way, models contribute to transparency and understanding. But model-based methodology can be applied to data that are not really random (as raised by Robert). Model checks cannot make sure that the models are 'true'; they help to avoid inappropriate and misleading analysis of the data. Romeijn's idea that true models will lead to true results, viewing statistics as logic, runs into problems here.

Methods that are not based on probability models perform well in some important tasks, but they come with their own implicit assumptions. We would argue that such methods work because they allow the flexible use of masses of data, which is often possible because the models are fitted by using regularization rather then pure optimization. Regularization in turn requires assumptions (or choices), which should be transparent and use subject-matter knowledge where possible: advice which we think is consistent with the experience of Murtagh on stable methods for 'big data' analytics.

One issue with supposedly assumption-free methodology is that it can discourage researchers from clarifying their specific research aims. In cluster analysis, for example, different research aims require different definitions of clusters, whereas model- and tuning-free methods are often advertised as making such decisions without 'subjective' researcher input; see Hennig (2015). That said, we welcome assumption-light or model-free theoretical guarantees as mentioned by Meilă. Lund and Iyer sketch a framework to analyse the interplay between data, assumptions and interpretation.

Several of the contributions challenge our attitude towards realism and truth (Marin and his colleagues, Paul and Romeijin), sometimes together with questioning consensus or stability as virtues (Robert, O'Rourke, Thall and VanderWeele). Realism and the meaning of truth have been controversial issues in philosophy for ages, and we shall not solve these controversies. 'How things really are', as O'Rourke cites Peirce, is ultimately inaccessible to human beings. We can do experiments, make observations and communicate them. The idea of some kind of objective, observer-independent reality ultimately relies on agreement about observations: in other words, some kind of consensus. Communication and consensus (and also stability and reproducibility; see Hannig's discussion) are crucial for making statements about reality that are claimed to hold for general observers. They are even crucial to set up the conditions for such statements such as agreed measurement procedures. 'Truth' can certainly be meaningfully used within systems of communication in which there is agreement about how to establish it, for example, within mathematics and logic or referring to measurements. The idea of truth applied to our models and many of our theories, though, leads us outside this domain. Betancourt and VanderWeele each take us on in a slightly different way, arguing that consensus should not be considered a virtue in itself but rather is—where it is appropriate at all—a consequence of other virtues. In contrast, Thall and O'Rourke remind us that a consensus is not so valuable if it is obtained too easily by following potentially misguided conventions. We still think of consensus as a valuable goal even though it can be abused in group decision making. Consensus does not always exist, but people are rightly troubled by nagging disagreements. Jukola's addition that institutional and social conditions of inquiry should be taken into account is valuable. Boulesteix and Strasser deal with the tension between multiple perspectives, stability and consensus, using ensemble approaches to bring multiple perspectives together, but also highlighting the danger that individual perspectives should not be driven by partiality.

Robert and Bandyopadhyay are concerned with the logical holes that are inherent in the 'falsificationist Bayesian' perspective. We are concerned about this also! From Gelman (2011);

> 'My point here is not to say that my preferred methods are better than others but rather to couple my admission of philosophical incoherence with a reminder that there is no available coherent alternative'.

We agree with Sprenger that it is only rarely that a Bayesian prior distribution completely mirrors a statistician's or researcher's beliefs; rather, priors and data generation models are imperfect codings of some subset of available information and thus are inherently subject to revision, and we agree with Morey that the practice or principle of parsimony fits awkwardly within our framework.

Moores points out that, following Lakatos rather than Popper, we are typically interested in improving rather than falsifying our models, which raises the question of what procedures should be used in the improvement step, considering that on Cantorian grounds we have already ruled out the existence of a supermodel that would allow model building, inference and improvement all to be done in one super-Bayesian step. Draper's approach using explicit conditioning is, we hope, a step in making our practices more logically and statistically coherent.

Maclaren asks whether we think of robustness in the 'Boxian' sense as a property of a good model that is flexible and not brittle, or in the 'Tukeyian' sense as a set of operations to be performed on data. We think that the two views are complementary, in that Boxian robust procedures can be considered as models that instantiate Tukeyian data trimming procedures—and vice versa.

Wynn argues that Bayesians should stop worrying about objectivity and instead should embrace context dependence and multiple perspectives (virtues V5 and V6) by considering the iterative process of model building, checking and improvement as occurring not inside the statistician's head but rather within a larger community of scientists and interested parties. We agree. But Celeux also has a good point when he says that the complexity of hierarchical Bayesian modelling makes transparency more difficult. Section 4.1 of our paper demonstrates how the specification of a prior distribution can be made more transparent by open admission of the process by which the prior has been constructed, but we agree that more work should be done in this area to make routine that sort of transparent exposition. Beyond this, as French notes, the most important aspects of a model are often 'structural' and encode substantive models or assumptions; we should have a way of explicating their sources also.

From a different Bayesian direction, Jewson argues that, in the real world 'M-open' scenario in which the true data-generating process is not included in the class of models that are used to fit the data, it could be possible to improve on straight Bayesian updating, thus introducing another element of researcher's choice into data analysis. We suppose that this could be incorporated in the usual Bayesian framework by instituting a loss function that captures which aspects of future data are of most interest. Relatedly, Leonelli and Smith connect the idea of 'institutional decision analysis' (see Section 3.1 of our paper) to their work on decision support systems, which represent another way to incorporate additional information and perspectives transparently into decision making.

Mateu goes even further, not just recommending context awareness but disagreeing with our characterization of statistics as a 'science of defaults' (Gelman, 2014). In response we, along with Vukcevic, Moreno-Betancur and Carlin, point to the world of practitioners in data collection and data analysis, many with little formal statistical training, who use available methods. Those of us who write software know that default settings are important, and much of our own research involves the development of methods that will work well the first time in a wide variety of realistic examples, and which are loud rather than quiet in their failure modes, to warn users away from some of the more disastrous consequences of inappropriate modelling and analysis choices. We appreciate that Wagenmakers likens our paper to Disneyland and we hope that he will forgive us our 'flirtation with frequentism' by recognizing the practical value for methodologists and textbook writers to gain some sense of the statistical properties of our methods—i.e. the consequences should our recommendations be followed in some specified distribution of cases.

Crane agrees with us that the terms 'objective' and 'subjective' have lost their value, but he would go one further than us by not attempting to categorize virtues at all. We agree with Crane that, ultimately, no principle is sacrosanct. For example one might think that correspondence with observed reality is essential to all science, but there can be value in purely theoretical work with no direct links to reality at all. Transparency is often beneficial but sometimes our methods outpace our understanding, and there can be value in black box methods that just seem to work: and so on. However, we do think that our list of virtues is useful, not as a checklist or requirement for published work but as an elaboration and exposition of the real concerns underlying all that talk about subjectivity and objectivity in statistics.

Vandemeulebroecke asks how our list of virtues would apply in confirmatory settings in drug development where 'emphasis is placed on type I error control and prespecification of analyses'. To the extent that researchers do care about such things, we would connect error control to impartiality (virtue V3) and prespecification to open planning and following agreed protocols (virtue V1(b)). The point here is not to match every statistical method to a virtue, but to understand that methods are devised and used to satisfy some goals—in this case, a desire to avoid making poor decisions based on noisy data. However, given that in reality effects are not exactly zero, we should be able to satisfy this aim in the context of a more realistic model of non-zero varying effects, which in turn will require some quantification of the costs and benefits of different decision options. Rather than setting a familywide error threshold of 0.05 or 0.01, implicitly motivated by some appeal to consensus (virtue V2(b)), we recommend that users determine decision thresholds based on real world contexts and aims (virtue V6(a)). This is an example of the larger point raised by Zyphur and Pierides, that our virtues can conflict with each other, but recognizing and adjudicating these conflicts can help us to understand the motivations for using different statistical rules better.

King points out that a systematic study of goals and virtues could be appropriate not just for probability modelling (the focus of our paper) but also when studying graphical communication and exploratory

data analysis. Bartholomew and Kumar each ask about extensions in a different way, to consider more formally the relationship between finite population inference, the statistical theory of sampling and real world surveys; our only thought here is that the relevant theory should address applied goals as well as the sampling process itself. And we agree with Vanpaemel, Tuerlinckx and De Boeck that the same principles occur for measurement and design as with data analysis: rather than hiding one's decisions under the rug of convention or presumed objectivity, we believe that it is better to be explicit about options and choices.

As Winkler explains, statisticians work at many levels: applied data collection and analysis, development of statistical theory and methods, construction of software and, not least, communication and teaching. The relevance of the virtues that we list for applied statistics may be somewhat different for those other tasks. Barker emphasizes that probability models are by their nature inexact, which motivates scepticism, a virtue that did not make it onto our list but which implicitly appears in many places, most notably openness to criticism (virtue V3(c)). Barker's concerns (a), (b) and (c) correspond to our virtues V1(c), V7(a) and V3(c).

Gepp, pointing to the literature on machine learning, discusses a way in which our virtues can interfere with each other, at least in the short term, so that direct pursuit of model improvement can reduce the correspondence of the fitted model to observed reality. In applied problems, there are many potential sources of information, and we think about the applied context when considering which variables to go to the trouble of measuring and modelling. Wijayatunga notes the importance of prediction accuracy as a criterion that should help with consensus—as long as there can be agreement on the outcomes to be predicted and the corpus of problems to average over. In a similar vein, Hannig explains how virtues such as transparency and stability can have different meanings under different inferential philosophies. Vukcevic, Moreno-Betancur and Carlin emphasize that the steps of statistical analysis (as well as design and data collection) are guided by the applied goals of any project. Statistics textbook writers tend to take goals for granted and then jump right into the modelling and data analysis without always specifying the connection between assumptions and goals.

Commenting on two of the more specific statistical issues, the notion of outliers (branded as 'nonsense in an M-open perspective' by Marin and his colleagues) was used in our presentation to illustrate how background information such as regarding to what extent the measurements in question are prone to gross errors has implications on data analysis. We think that outliers are an important concept in the interpretation of data in many applications, regardless of what methodology and models are applied, Montanari asks about the connection between non-identifiability and multiple perspectives. We think it important with non-identified or weakly identified models to acknowledge uncertainty, so that, if one solution presented is mathematically equivalent to many others, the results cannot be distinguished, and it would be inappropriate to pick one solution and to sell it as the unique answer. However, it is legitimate to present the result as compatible with multiple perspectives, and to select one for being, for example, the easiest possible interpretation.

Ultimately we agree with VanderWeele and Paul regarding the goal of a 'consensus-based, rigorously vetted, scientific representation of the worlds'. Much of the tension in our paper, the discussants and decades of earlier commentary on objectivity and subjectivity in statistics comes from the goal of using flawed models and approximate methods to solve real problems and to learn general truths. We hope that a better understanding of 'the unreasonable effectiveness of statistics' (to paraphrase Wigner (1960)) will give us insights into how to make these tools even more effective.

References in the discussion

Ackrill, J. L. (1963) *Categories and De Interpretatione* (Engl. transl.). New York: Oxford University Press.

Akaike, H. (1973) Information theory as an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Information Theory* (eds B. N. Petrov and C. Csàki), pp. 267–281. Budapest: Akademiai Kiado.

Bandyopadhyay, P. (2007) Why Bayesianism?: A primer on a probabilistic philosophy of science. In *Bayesian Statistics and Its Applications* (eds S. Upadhyay, U. Sing and D. Dey), pp. 42–62. New Delhi: Anamaya.

Bandyopadhyay, P., Bennett, J. and Higgs, M. (2015) How to undermine underdetermination? *Foundns Sci.*, **20**, 107.

Bandyopadhyay, P. and Boik, R. (1999) The curve-fitting problem: a Bayesian rejoinder. *Phil. Sci.*, **66**, suppl., S390–S402.

Bandyopadhyay, P., Boik, R. and Basu, P. (1996) The curve-fitting problem: a Bayesian approach. *Phil. Sci.*, **63**, suppl., S264–S272.

Bandyopadhyay, P. and Brittan, G. (2002) Logical consequence and beyond: a look at model selection in statistics. In *Logical Consequence and Beyond* (eds J. Woods and B. Hepburn), pp. 1–14. Oxford: Hermes.

Bandyopadhyay, P., Brittan, G. and Taper, M. (2016) *Belief, Evidence, and Uncertainty: Problems of Epistemic Inference*. New York: Springer.
Bandyopadhyay, P. and Forster, M. (eds) (2011) *Handbook of Philosophy of Statistics*. Amsterdam: Elsevier.
Bandyopadhyay, P., Greenwood, M. and Brittan, G. (2014) Empiricism and/or instrumentalism. *Erkenntnis*, **79**, suppl. 5, 1015.
Bandyopadhyay, P., Taper, M. and Britain, G. (2017) Non-Bayesian account of evidence: Howson's counter-example countered. *Int. Stud. Phil. Sci.*, to be published.
Barad, K. (1996) Meeting the universe halfway: realism and social constructivism without contradiction. In *Feminism, Science, and the Philosophy of Science* (eds L. Nelson and J. Nelson), pp. 161–194. Berlin: Springer.
Barad, K. (1998) Getting real: Technoscientific practices and the materialization of reality. *Differences*, **10**, 87–91.
Barad, K. (2007) *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Durham: Duke University Press.
Bartholomew, D. J., Steel, F., Moustaki, I. and Galbraith, J. I. (2008) *Analysis of Multivariate Social Science Data*. Boca Raton: Chapman and Hall–CRC.
Benzécri, J. P. (1983) L'avenir de l'analyse des données (The future of data analysis). *Behaviormetrika*, **10**, 1–11.
Berger, J. O., Bernardo, J. M. and Sun, D. (2009) The formal definition of reference priors. *Ann. Statist.*, **37**, 905–938.
Bernardo, J. M. and Smith, A. F. M. (2000) *Bayesian Theory*. Chichester: Wiley.
Bissiri, P. G., Holmes, C. C. and Walker, S. G. (2016) A general framework for updating belief distributions. *J. R. Statist. Soc.* B, **78**, 1103–1130.
Boulesteix, A.-L., Hable, R., Lauer, S. and Eugster, M. J. (2015) A statistical framework for hypothesis testing in real data comparison studies. *Am. Statistn*, **69**, 201–212.
Boulesteix, A.-L., Hornung, R. and Sauerbrei, W. (2017) On fishing for significance and statistician's degree of freedom in the era of big molecular data. In *Berechenbarkeit der Welt?: Philosophie und Wissenschaft in Zeitalter von Big Data* (eds J. Wernecke, W. Pietsch and M. Otte). Berlin: Springer.
Boulesteix, A.-L., Lauer, S. and Eugster, M. J. A. (2013) A plea for neutral comparison studies in computational sciences. *PLOS ONE*, **8**, article e61562.
Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations (with discussion). *J. R. Statist. Soc.* B, **26**, 211–252.
Bro, R. and Smilde, A. K. (2003) Centering and scaling in component analysis. *J. Chemometr.*, **17**, 16–33.
Button, K. S., Ionnidis, J. P. A., Mokrysz, C., Nosek, B. A. Flint, J., Robinson, E. S. J. and Munafo, M. R. (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neursci.*, **14**, 365–376.
Callahan, B., Proctor, D., Relman, D., Fukuyama, J. and Holmes, S. (2016) Reproducible research workflow in R for the analysis of personalized human microbiome data. *Pac. Symp. Biocomput.*, **21**, 183–194.
Cappelen, H. and Dever, J. (2016) *Context and Communication (Contemporary Introductions to Philosophy of Language)*. Oxford: Oxford University Press.
Chen, C.-L., Gilbert, T. J. and Daling, J. (1999) Maternal smoking and Down syndrome: the confounding effect of maternal age. *Am. J. Epidem.*, **149**, 442–446.
Christie, M., Cliffe, A., Dawid, A. P. and Senn, S. (eds) (2011) *Simplicity, Complexity and Modelling*. Chichester: Wiley.
Cooke, R. M. (1991) *Experts in Uncertainty*. Oxford: Oxford University Press.
Cox, N. J. (2013) Trimming to taste. *Stata J.*, **13**, 640–666.
Davies, P. L. (2014) *Data Analysis and Approximate Models*. Boca Raton: CRC Press.
Davison, A. C. and Hinkley, D. V. (1997) *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.
Dawid, A. P. (1982) The well-calibrated Bayesian (with discussion). *J. Am. Statist. Ass.*, **77**, 605–613.
Dawid, A. P. (1984) Statistical theory: the prequential approach (with discussion). *J. R. Statist. Soc.* A, **147**, 278–292.
Dawid, A. P. (1985) Calibration-based empirical probability (with discussion). *Ann. Statist.*, **13**, 1251–1285.
Dawid, A. P. (2004) Probability, causality and the empirical world: a Bayes–de Finetti–Popper–Borel synthesis. *Statist. Sci.*, **19**, 44–57.
Dawid, A. P. (2010) Beware of the DAG! *J. Mach. Learn. Res. Wrkshp Conf. Proc.*, **6**, 59–86.
DeRose, K. (2009) *The Case of Contextualism*. Oxford: Oxford University Press.
Dewey, J. (1938) *Logic: the Theory of Inquiry*. New York: Holt.
Ding, P. and VanderWeele, T. J. (2016) Sensitivity analysis without assumptions. *Epidemiology*, **27**, 368–377.
Douglas, H. (2000) Inductive risk and values in science. *Phil. Sci.*, **67**, 559–579.
Douglas, H. (2004) The irreducible complexity of objectivity. *Synthese*, **138**, 453–473.
Douglas, H. (2009) *Science, Policy, and the Value-free Ideal*. Pittsburgh: Pittsburgh University Press.
Draper, D. (2013) Bayesian model specification: heuristics and examples. In *Bayesian Theory and Applications* (eds P. Damien, P. Dellaportas, N. Polson and D. Stephens), ch. 20, pp. 409–431. Oxford: Oxford University Press.
Evans, M. (2015) *Measuring Statistical Evidence using Relative Belief*. Boca Raton: Chapman and Hall–CRC.
Feinstein, L. (2003) Inequality in the early cognitive development of British children in the 1970 cohort. *Economica*, **70**, 73–97.

Feyerabend, P. (2010) *Against Method*, 4th edn. London: Verso.
French, S. (2012) Expert judgment, meta-analysis, and participatory risk analysis. *Decsn Anal.*, **9**, 119–127.
Friedman, J. H. (1997) On bias, variance, 0/1-loss and the curse-of-dimensionality. *Data Mining Knowl. Discov.*, **1**, 55–77.
Galbraith, J. I. and Stone, M. (2011) The abuse of regression in the National Health Service allocation formulae: response to the Department of Health's 2007 'resource allocation research paper' (with discussion). *J. R. Statist. Soc.* A, **174**, 517–528; 547–567.
Gelman, A. (2003) A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *Int. Statist. Rev.*, **71**, 369–382.
Gelman, A. (2007) Bayesian checking of the second levels of hierarchial models. *Statist. Sci.*, **22**, 349–352.
Gelman, A. (2011) Induction and deduction in Bayesian data analysis. *Rationalty, Markts Morals*, **2**.
Gelman, A. (2014) How do we choose our default methods? In *Past, Present, and Future of Statistical Science* (eds X. Lin, C. Genest, D. L. Banks, G. Molenberghs, D. W. Scott and J. L. Wang), pp. 293–301. London: Chapman and Hall.
Gelman, A. (2017) Honesty and transparency are not enough. *Chance*, **30**, 37–39.
Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013) *Bayesian Data Analysis*, 3rd edn. Boca Raton: Chapman and Hall–CRC.
Gelman, A. and Loken, E. (2014) The statistical crisis in science. *Am. Scient.*, **102**, 460–465.
Gelman, A. and Shalizi, C. R. (2013) Philosophy and the practice of Bayesian statistics. *Br. J. Math. Statist., Psychol.*, **66**, 8–38.
Gigerenzer, G. and Marewski, J. N. (2015) Surrogate science: the idol of a universal method for scientific inference. *J. Mangmnt*, **41**, 421–440.
Good, I. J. (1950) *Probability and the Weighing of Evidence*. London: Griffin.
Good, I. J. (1983) *Good Thinking: the Foundations of Probability and Its Applications*. Minneapolis: University of Minnesota Press.
Good, I. J. (1989) *Good Thinking: the Foundations of Probability and Its Applications*, 2nd edn. Mineola: Dover Publications.
Goodman, S. N. (2016) Aligning statistical and scientific reasoning. *Science*, **352**, 1180–1181.
Greenland, S. (2005) Multiple-bias modelling for analysis of observational data (with discussion). *J. R. Statist. Soc.* A, **168**, 267–306.
Gregor, M. (2015) *Critique of Practical Reason* (Engl. transl.), 2nd edn. Cambridge: Cambridge University Press.
Hacking, I. (2015) Let's not talk about objectivity. In *Objectivity in Science*, pp. 19–33. New York: Springer.
Hand, D. (2009) Mismatched models, wrong results, and dreadful decisions: on choosing appropriate data mining tools. In *Proc. 15th Int. Conf. Knowledge Discovery and Data Mining*, pp. 1–2. Paris: Association for Computing Machinery.
Hannig, J., Iyer, H., Lai, R. C. S. and Lee, T. C. M. (2016) Generalized fiducial inference: a review and new results. *J. Am. Statist. Ass.*, **111**, 1346–1361.
Harding, S. (1991) *Whose Science?; Whose knowledge?: Thinking from Women's Lives*. Ithaca: Cornell University Press.
Hempel, C. G. (1965) *Aspects of Scientific Explanation*. New York: Free Press.
Hennig, C. (2010) Mathematical models and reality: a constructivist perspective. *Foundns Sci.*, **15**, 29–48.
Hennig, C. (2015) What are the true clusters? *Pattn Recogn Lett.*, **64**, 53–62.
Hicks, D. J. and Stapleford, T. A. (2016) The virtues of scientific practice: MacIntyre, virtue ethics, and the historiography of science. *Isis*, **107**, 449–472.
Höhle, U., (1990) The Poincaré paradox and the cluster problem. In *Trees and Hierarchical Structures* (eds A. Dress and A. von Haeseler). Berlin: Springer.
Huber, P. J. and Ronchetti, E. M. (2009) *Robust Statistics*, 2nd edn. Hoboken: Wiley.
Intergovernmental Panel on Climate Change (2013) The physical science basis: contribution of Working Group I. In *The Fifth Assessment Report of Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.
Jaynes, E. (2003) *Probability Theory*. Cambridge: Cambridge University Press.
Jeffreys, H. (1961) *Theory of Probability*. London: Oxford University Press.
Jeffreys, H. (1963) Review of "The Foundation of Statistical Inference". *Technometrics*, **3**, 407–410.
Jukola, S. (2017) On ideal of objectivity, judgements and bias in medical research—a comment on Stegenga. *Stud. Hist. Phil. Sci.* C, **62**, 35–41.
Klein, R. A., Ratliff, K. A., Vianello, Jr, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., Hasselman, F., Hicks, J. A., Hovermale, J. F., Hunt, S. J., Huntsinger, J. R., Uzerman, H., John, M.-S., Joy-Gaba, J. A., Kappes, H. B., Krueger, L. E., Kurtz, J., Levitan, C. A., Mallett, R. K., Morris, W. L., Nelson, A. J., Nier, J. A., Packard, G., Pilati, R., Rutchick, A. M., Schmidt, K., Skorinko, J. L., Smith, R., Steiner, T. G., Storbeck, J., Swol, L. M. V., Thompson, D., van 't Veer, A. E., Vaughn, L. A., Varanka, M., Wichman, A. L., Woodzicka, J. A. and Nosek, B. A. (2014) Investigating variation in replicability. *Socl Psychol.*, **45**, 142–152.

Kuhn, T. (1962) *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
van der Laan, M. J. and Rose, S. (2011) *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer.
Lash, T. L., Fox, M. P. and Fink, A. K. (2009) *Applying Quantitative Bias Anaylsis to Epidemiologic Data*. New York: Springer.
Leonelli, M. and Smith, J. Q. (2015) Bayesian descision support for complex systems with many distributed experts. *Ann. Oper. Res.*, **235**, 517–542.
Lipton, P. (2004) *Inference to the Best Explanation*, 2nd edn. Abingdon: Routledge.
Liu, K. and Meng, X.-L. (2016) There is individualized treatment: why not individualized inference? *A. Rev. Statist. Appl.*, **3**, 79–111.
Lloyd, S. P. (1982) Least squares quantization in PCM. *IEEE Trans. Inform. Theory*, **28**, 129–137.
Lonergan, B. J. F. (1965) *Insight: a Study of Human Understanding*. New York: Philosophical Library.
Longford, N. T. (2012) 'Which model?' is the wrong question. *Statist. Neerland.*, **66**, 237–252.
Longford, N. T. (2013) *Statistical Decision Theory*. Heidelberg: Springer.
Longford, N. T. (2016) Comparing two treatments by decision theory. *Pharm. Statist.*, **15**, 387–395.
Longino, H. (1990) *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton: Princeton University Press.
Lord, F. M. (1967) A paradox in the interpretation of group comparisons. *Psychol. Bull.*, **68**, 304–305.
Lund, S. P. and Iyer, H. K. (2017) Likelihood ratio as weight of forensic evidence: a closer look. *J. Res. Natn. Bur. Stand. Technol.*, to be published.
Mayo, D. G. (1996) *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.
Meilă, M. (2006) The uniqueness of a good optimum for K-means. In *Proc. Int. Machine Learning Conf*. (eds A. Moore and W. Cohen), pp. 625–632. New York: International Machine Learning Society.
Murtagh, F. (2016) Sparse p-adic data coding for computationally efficient and effective Big Data analytics. *Ultrametr. Anal. Appl.*, **8**, 236–247.
Murtagh, F. (2017) *Data Science Foundations: Geometry and Topology of Complex Hierarchic Systems and Big Data Analytics*. Boca Raton: Chapman and Hall–CRC.
Nuzzo, R. (2014) Scientific method: statistical errors. *Nature*, **506**, 150–152.
Oakley, J. E. and O'Hagan, A. (2004). Probabilistic sensitivity analysis of complex models: a Bayesian approach. *J. R. Statist. Soc.* B, **66**, 751–769.
Pages, J. (2015) *Multiple Factor Analysis by Example Using R*, p. 283. Boca Raton: Chapman and Hall.
Paul, L. A. (2014) *Transformative Experience*. Oxford: Oxford University Press.
Paul, L. A. and Healy, K. (2017) Transformative treatments. *Noûs*, to be published.
Pearl, J. (2009) *Causality: Models, Reasoning, and Inference*, 2nd edn. Cambridge: Cambridge University Press.
Pearl, J. (2016) Lord's paradox revisited—(Oh Lord Kumbaya!). *J. Causl Inf.*, **4**, no. 2.
Peirce, C. S. (1878) Deduction, induction, and hypothesis. *Poplr Sci. Mnthly*, **13**, 470–482.
Peirce Edition Project (eds) (1998) *The Essential Peirce: Selected Philosophical Writings (1893–1913)*. Bloomington: Indiana University Press.
Poovey, M. (1998) *A History of the Modern Fact: Problems of Knowledge in the Sciences of Wealth and Society*. Chicago: University of Chicago Press.
Popper, K. R. (1979) *Objective Knowledge: an Evolutionary Approach*. Oxford: Oxford University Press.
Porter, T. M. (1991) Objectivity and authority: how French engineers reduced public utility to numbers. *Poet. Today*, **12**, 245–265.
Porter, T. M. (1992) Objectivity as standardization: the rhetoric of impersonality in measurement, statistics, and cost-benefit analysis. *Ann. Scholshp*, **9**, 19–59.
Porter, T. M. (1993) Statistics and the politics of objectivity. *Rev. Synth.*, **114**, 87–101.
Porter, T. M. (1995) *Trust in Numbers: the Pursuit of Objectivity in Science and Public Life*. Princeton: Princeton University Press.
Reid, W. V., Chen, D., Goldfarb, L., Hackmann, H., Lee, Y. T., Mokhele, K., Ostrom, E., Raivio, K., Rockström, J., Schellnhuber, H. J. and Whyte, A. (2010) Earth system science for global sustainability: grand challenges. *Science*, **330**, 916–917.
Reiss, J. and Sprenger, J. (2014) Scientific objectivity. In *The Stanford Encyclopedia of Philosophy* (ed. E. N. Zalta). Stanford: Stanford University.
Rosenberger, J. L. and Gasko, M. (1983) Comparing location estimators: trimmed means, medians, and trimean. In *Understanding Robust and Exploratory Data Analysis* (eds D. C. Hoaglin, F. Mosteller and J. W. Tukey), pp. 297–338. New York: Wiley.
Royall, R. (1997) *Statistical Evidence*. New York: Chapman and Hall.
Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, **66**, 688–701.
Rubin, D. B. (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, **12**, 1151–1172.
Rudner, R. (1953) The scientist *qua* scientist makes value judgments. *Phil. Sci.*, **30**, 1–6.
Sauerbrei, W., Abrahamowicz, M., Altman, D. G., le Cessie, S. and Carpenter, J. (2014) STRengthening Analytical Thinking for Observational Studies: the STRATOS initiative. *Statist. Med.*, **33**, 5413–5432.

Seillier-Moiseiwitsch, F. and Dawid, A. P. (1993) On testing the validity of sequential probability forecasts. *J. Am. Statist. Ass.*, **88**, 355–359.
Seillier-Moiseiwitsch, F., Sweeting, T. J. and Dawid, A. P. (1992) Prequential tests of model fit. *Scand. J. Statist.*, **19**, 45–60.
Senn, S. (2003) Disappointing dichotomies. *Pharm. Statist.*, **2**, 239–240.
Shapin, S. (1994) *A Social History of Truth: Civility and Science in Seventeenth-century England*. Chicago: University of Chicago Press.
Shapin, S. (2008) *The Scientific Life: a Moral History of a Late Modern Vocation*. Chicago: University of Chicago Press.
Shapin, S. (2016) A taste of science: making the subjective objective in the California wine world. *Socl Stud. Sci.*, **46**, 436–460.
Sheiner, L. B. (1997) Learning versus confirming in clinical drug development. *Clin. Pharmcol. Ther.*, **61**, 275–291.
Silver, N. (2012) *The Signal and the Noise: the Art and Science of Prediction*. New York: Penguin.
Simmons, J., Nelson, L. and Simonsohn, U. (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allow presenting anything as significant. *Psychol. Sci.*, **22**, 1359–1366.
Smith, J. Q., Barons, M. J. and Leonelli, M. (2015) Coherent frameworks for statistical inference serving integrating decision support systems. *Research Report 15-09*. Department of Statistics, University of Warwick, Coventry.
Sprenger, J. (2017a) The objectivity of subjective Bayesianism. *Manuscript*. Tilburg University, Tilburg. (Available from `http://philsci-archive.pitt.edu/13199`.)
Sprenger, J. (2017b) Conditional degree of belief. *Manuscript*. Tilburg University, Tilburg. (Available from `http://philsci-archive.pitt.edu/12304`.)
Stehlík, M., Aguirre, P., Girard, S., Jordanova, P., Kiselák, J., Torres Leiva, S., Sadovský, Z. and Rivera, A. (2017) On ecosystems dynamics. *Ecol. Complex.* C, **29**, 10–29.
Stehlík, M., Dušek, J. and Kiselák, J. (2016) Missing chaos in global climate change data interpreting? *Ecol. Complex.*, **25**, 53–59.
Stehlík, M., Helpersdorfer, Ch., Hermann, P., Šupina, J., Grilo, L. M., Maidana, J. P., Fuders, F. and Stehlíková, S. (2017) Financial and risk modelling with semicontinuous covariances. *Inform. Sci.* C, **394–395**, 246–272.
Stigler, S. M. (2016) *The Seven Pillars of Statistical Wisdom*. Boston: Harvard University Press.
Suarez, M. (2017) Propensities, probabilities and experimental statistics. In *Proc. 5th Conf. European Philosophy of Science Association* (eds M. Massimi and J. W. Romeijn). New York: Springer.
Tufte, E. R. (1983) *The Visual Display of Quantitative Information*. Storrs: Graphics Press.
Tukey, J. W. (1962) The future of data analysis. *Ann. Math. Statist.*, **33**, 1–67.
Tukey, J. W. (1977) *Exploratory Data Analysis*. Reading: Addison-Wesley.
Urbanek, S. (2006) *Exploratory Model Analysis: an Interactive Graphical Framework for Model Comparison and Selection*. Norderstedt: Books on Demand.
Vapnik, V. (1998) *Statistical Learning Theory*. New York: Wiley.
Wasserman, L. (2012) A world without referees. Carnegie Mellon University, Pittsburgh. (Available from `http://www.stat.cmu.edu/larry/Peer-Review.pdf`.)
Wasserstein, R. L. and Lazar, N. A. (2016) The ASA's statement on p-values: context, process, and purpose. *Am. Statistn*, **70**, 129–133.
Wickham, H. (2006) Exploratory model analysis with R and GGobi. *Technical Report*. (Available from `http://had.co.nz/model-vis/2007-jsm.pdf`.)
Wickham, H. (2017) The tidyverse. (Available from `http://tidyverse.org`.)
Wigner, E. (1960) The unreasonable effectiveness of mathematics in the natural sciences. *Communs Pure Appl. Math.*, **13**, 1–13.
Wilkinson, L. (2005) *The Grammar of Graphics*, 2nd edn. New York: Springer.
Williams, N. and Stone, M. (2013) Explicating 'wrong' or questionable signs in England's NHS funding formulas: correcting wrong explanations. *Report*. Civitas, London.
Wolpert, D. H. (1996) The lack of *a priori* distinctions between learning algorithms. *Neurl Comput.*, **8**, 1341–1390.