Economics and Philosophy, page 1 of 30
doi:10.1017/S0266267118000159
cambridge.org/eap

TRADE-OFFS BETWEEN EPISTEMIC AND MORAL VALUES IN EVIDENCE-BASED POLICY

DONAL KHOSROWI*

Abstract: Proponents of evidence-based policy (EBP) call for public policy to be informed by high-quality evidence from randomized controlled trials. This methodological preference aims to promote several epistemic values, e.g. rigour, unbiasedness, precision, and the ability to obtain causal conclusions. I argue that there is a trade-off between these epistemic values and several non-epistemic, moral and political values. This is because the evidence afforded by standard EBP methods is differentially useful for pursuing different moral and political values. I expand on how this challenges ideals of value-freedom and -neutrality in EBP, and offer suggestions for how EBP methodology might be revised.

Keywords: evidence-based policy, values, trade-off, value-freedom, value-neutrality

1. INTRODUCTION

Evidence-based policy (EBP) is the movement according to which public policy should be informed by high-quality empirical evidence for policy effectiveness from randomized controlled trials (RCTs) and meta-analyses. EBP advocates' emphasis on the superior epistemic credentials of these methods can be understood to derive from several epistemic

* Department of Philosophy, Durham University, 50 Old Elvet, Durham DH1 3HN, UK
Email: donal.khosrowi@durham.ac.uk

values such as methodological rigour, unbiasedness, precision, and the ability to obtain causal conclusions about policy effectiveness.

In what follows I argue that these epistemic values stand in a *trade-off relation* with a wide range of moral values underlying the policy goals that policymakers may be interested in achieving with the help of effectiveness evidence. Specifically, I argue that standard EBP methodological recommendations, and the evidence produced in accordance with them, can severely inhibit policymakers' ability to pursue moral values such as equality or priority for the worst-off on the basis of typical EBP research outputs. This is because standard EBP methods are not informative about the *distributive consequences* of policies (see e.g. Manski 2000; Deaton 2010 for similar concerns). This is a substantive shortcoming, particularly when there are reasons to suspect that a policy intervention may make some individuals worse off. Since the evidence typically afforded by EBP methods is uninformative about such distributive consequences, it is *differentially useful* for the pursuit of different moral and political values, specifically values that do or do not put emphasis on how benefits and harms induced by policies are distributed among individuals. I argue that this differential usefulness gives rise to a trade-off between epistemic and non-epistemic moral values, where, relative to current EPB methodological tenets, advances to produce more informative evidence on distributive effects come at the expense of sacrificing several key EBP epistemic values at once. I elaborate how this may challenge both value-freedom and neutrality in EBP.

The contents are organized as follows. In Section 2 I reconstruct some of the key epistemic values involved in EBP as well as whether and how EBP involves ideals of value-freedom and neutrality. In Section 3 I expand on the epistemic challenges that EBP methodology faces with respect to generating information about the distributive consequences of policy interventions. In Section 4 I offer my argument for the trade-off between central EBP epistemic values and moral values that are sensitive to distributive consequences of policies. I also expand on the consequences for value-freedom and neutrality in EBP and elaborate how an epistemic-ethical approach might be a promising way of facilitating deliberation about how this situation should be addressed. Section 5 concludes.

2. VALUES IN EBP

To build a basis for developing my subsequent arguments, let me begin with two short reconstructions. The first concerns the central epistemic values that I take to be underlying EBP methodology. The second concerns what I consider to be the roles of value-freedom and -neutrality in EBP respectively.

2.1. Epistemic Values in EBP

It is important to note that there is perhaps no universally accepted set of epistemic values common to all activities under the EBP heading. More fundamentally, it may be contested whether there is something like a unified EBP paradigm at all. The EBP movement, particularly as it changes over time and in response to various criticisms, is difficult to precisely demarcate as a unified paradigm with distinctive and invariant objectives, methods, underlying epistemic value presuppositions and so forth.[1] Moreover, it is also difficult to find clear-cut commitments to particular sets of epistemic values, which means that the reconstruction I offer below will be just that: a reconstruction that is potentially imperfect and incomplete in its rendition of the actual roles of epistemic values in EBP.

That being said, it seems uncontroversial that there is a kernel of epistemic values that are common to a broad variety of activities under the EBP heading. This is because many of these activities are conducted in accordance with standard methodological recommendations offered by key institutions, such as the Campbell Collaboration, the What Works Clearinghouse, CONSORT, GRADE, the Cochrane Collaboration, JPAL and others. Many of these methodological recommendations, in particular so-called quality-of-evidence ranking schemes that rank different kinds of evidence according to quality and credibility, overlap in their emphasis on a kernel of epistemic values. These values are not coextensive with traditional epistemic values in the context of theory choice or appraisal such as those offered by Kuhn (1977). Instead, for empirical paradigms such as EBP the core epistemic values of interest concern the estimation of policy effects and include the *methodological rigour* that should be exercised when using different methods to obtain such estimates; the *unbiasedness* and *precision* of these estimates; and *the ability to obtain causal conclusions* about policy interventions on grounds of such estimates.

More specifically, *methodological rigour* is broadly understood as an attitude towards focusing on the use of only those methods that are believed to be (most) reliable in producing correct estimates of the quantities one is interested in measuring; the thorough exercise of appropriate precautions to ensure and demonstrate that the assumptions

[1] In addition to this caveat, it is important to note that the construal of Evidence-Based Policy I consider here is somewhat narrow in that it focuses on the so-called *treatment effects literature* as instantiated in e.g. development economics and educational research. The distinctive characteristic of this literature is its predominant focus on experimental and quasi-experimental methods to estimate treatment effects. This is considerably narrower than a construal of evidence-based policy as policy that is informed by any empirical evidence rather than only specific kinds of such evidence. I thank Erin Nash for calling for clarification.

required for the successful use of these methods are valid; as well as a clear preference for methods that require few substantive assumptions to begin with.

Unbiasedness is typically understood in the sense that the effects estimated by using certain methods should identify only the quantities that one is interested in measuring, e.g. the causal effect of an intervention on X on an outcome Y, rather than, for instance, also capturing the effects of other variables on Y.

Precision is the requirement that studies should be adequately powered to detect the effects of interest, e.g. by ensuring that sample sizes are sufficiently large and that error bounds on the estimates of interest (although not always computable without substantive assumptions, see e.g. Deaton 2010) are sufficiently small to minimize uncertainties about the magnitudes and signs of the effects being estimated.

Finally, the *ability to obtain causal conclusions* about policy effectiveness is the central idea that EBP research should focus on the production of quantities that are actionable for the purposes of policymakers, i.e. quantities, which, if correctly estimated, can figure as a basis for designing successful policy interventions. For that, the estimated effects must be successfully *causally* attributed to the intervention variables of interest, as otherwise subsequent intervention on such variables may fail to produce the expected or desired effects.

The above values seem central to EBP in the sense that they seem to underpin many aspects of standard EBP methodology, i.e. a set of salient methodological principles that are shared among proponents of the paradigm and are widely circulated in methodological recommendations, guidelines and manuals that advise practitioners on how to conduct and evaluate studies on policy effectiveness. For instance, EBP methodological recommendations specifically focus on certain epistemic targets, i.e. causal conclusions about policy effectiveness that are informative for policy formation. Moreover, EBP methodology is premised on principles concerning the relative desirability of certain kinds of evidence, e.g. by using quality-of-evidence ranking schemes to express a strict preference for experimental and quasi-experimental over purely observational evidence. Finally, EBP methodology emphasizes the relative ability of different methods with respect to generating desirable kinds of evidence; again by focusing on RCTs and quasi-experimental designs as opposed to observational studies. Together, these methodological tenets mediate between epistemic values and method choice in the sense that EBP methodological recommendations, such as those issued by quality-of-evidence ranking schemes, seem to advocate the use of certain methods, notably RCTs, specifically because these methods are considered to best promote the achievement of crucial epistemic values such as those outlined above.

With this brief reconstruction of key EBP epistemic values in place, let me turn to reconstruct the role of value-freedom and -neutrality ideals in EBP.

2.2. Value-Freedom and -Neutrality in EBP

It is not obvious that EBP proponents in general pursue any specific ideal with respect to value-freedom and -neutrality, i.e. concerning whether non-epistemic, moral values may play a role in the conduct of EBP research, and whether the outputs of this research may involve substantive moral value presuppositions. This is particularly difficult to tell as EBP methodological guidelines rarely comment on value-related issues.

Despite the difficulties involved in finding explicit commitments to ideals of value-freedom and -neutrality, it seems plausible to think that the EBP movement rests on a relatively broad axiological presupposition that some *division of labour* with regard to settling normative issues of what values public policy should promote, and settling factual issues of what are effective interventions to promote these values, is possible. In other words, while it is clear that evidence-based *policy* invariably involves moral and political values when it comes to specifying what outcomes policies should promote, EBP seems to rest on the assumption that agreement on the desirability of policy outcomes can be *separated* from the production of evidence speaking for the effectiveness of interventions in realizing these outcomes. On this view, whether or not a policy intervention such as reducing class sizes is effective in increasing student performance is an issue that can (and should) be settled independently of whether it is in fact desirable to increase student performance. The latter is supposed to be a question of values, the former a question of empirical facts pertaining to 'what works'; and it is the focus on documenting 'what works' that EBP is expressly committed to.

This broadly parallels traditional ideals regarding the role of non-epistemic, moral values in economics, where economists have frequently invoked the metaphor of *economists as social engineers*, who provide factual answers to policy questions *independently from* and typically after policymakers have settled issues concerning the relative desirability of social outcomes (cf. Hausman and McPherson 1996). While EBP proponents may not subscribe to this particular ideal, EBP methodology seems to presuppose at least that *some such* division of labour is possible in the sense that one can empirically investigate the effectiveness of policy interventions largely independently from questions concerning the desirability of the outcomes that they bring about (see Vedung 2010 for a similar reconstruction). Let me expand on what this suggests for the role of value-freedom and -neutrality in EBP.

First, it seems that EBP involves some ideal of value-freedom in the sense that non-epistemic values are generally not and should not be involved in shaping the conduct and outcomes of EBP research *internally*. According to such an ideal, while non-epistemic values may be involved in selecting the kinds of policy issues and interventions being studied, the outcome variables of interest, and may act as constraints on whether conducting RCTs is morally permissible, non-epistemic, moral values are generally not and should not be involved in the choice and application of methods once these issues are settled (see Biddle 2013: 124 for a related sketch of such a demarcation). For instance, the choice between RCTs and observational studies, the collection of data for such studies, or the interpretation of estimates obtained from such studies, should not vary with respect to researchers' (or policymakers') convictions concerning the desirability of the policies under scrutiny. These *internal* aspects should be guided by epistemic values alone.

Second, EBP seems to involve *some* version of value-neutrality in the sense that the outcomes of EBP research are intended to be value-neutral insofar as they should not, and generally do not, issue unconditional normative claims regarding the desirability of social outcomes. At most, *if* there are normative claims issued in the dissemination of EBP research, these claims take the shape of *hypothetical imperatives*, i.e. normative claims that are conditional on some substantive value presupposition but do not endorse this value presupposition as such. In order for EBP research to maintain value-neutrality, the adequacy of presuppositions speaking for the desirability of some social outcome must be settled *independently* from (and perhaps prior to) generating information about the effectiveness of different interventions in producing the outcome. If such independence is achieved, then *even if* EBP research sometimes issues normative claims, e.g. when researchers suggest that some intervention is preferable to others because it is more (cost-) effective, these claims would still be value-neutral since they remain non-committal on the adequacy of the substantive moral value presuppositions involved, i.e. whether the outcomes of interest are *in fact* morally desirable. These issues are left to policymakers and otherwise suitably legitimized agents to settle.

With this brief reconstruction of the central epistemic values in EBP as well as the role of value-freedom and -neutrality in place, let me focus on the main issue that I am interested in, which is that standard EBP methods are differentially suitable for the pursuit of different kinds of moral and political values, specifically values that do or do not put emphasis on the distributive consequences of policies. This differential usefulness creates a trade-off relation between epistemic and moral values in EBP. To explain why this is so, let me begin by offering some background on the epistemic challenges involved in producing evidence on the distributive consequences of policy interventions.

3. TREATMENT EFFECT HETEROGENEITY

Public policy interventions almost invariably affect agents in heterogeneous ways. Consider for instance the case of *microfinance programmes*, i.e. programmes that supply microcredit to agents who lack access to capital markets. Let us grant for the moment that at least some of these programmes may be successful in generating some kinds of positive long-term consequences, e.g. by increasing average household endowment or private investment in durable goods. Even so, behavioural response to microfinance access often differs significantly between agents (see e.g. Banerjee *et al*. 2017). Some agents, e.g. those whose otherwise successful entrepreneurial efforts are inhibited by inadequate access to capital markets, may significantly benefit from such programmes. Yet, other, economically less sophisticated agents may be driven into debt traps by pursuing unprofitable business plans and taking up high-interest loans in order to repay initial programme loans.

Such heterogeneity in individual treatment effects is attributable to differences in the causal mechanisms involved in the production of the outcomes of interest or the individual-specific realizations of variables that figure in these mechanisms. Specifically, the causal mechanisms connecting treatment and outcome variables typically involve what I will call *interactive covariates* of the treatment effect, i.e. variables that causally interact with the treatment and can modify the magnitude and/or sign of the causal effect induced by one and the same intervention.[2] For instance, the mechanisms that causally relate microfinance access and eventual welfare outcomes of agents plausibly involve an extensive battery of interactive covariates such as entrepreneurial ability, education, prior business ownership, pre-intervention budget constraints, business plan feasibility etc. As individuals will typically differ in their individual-specific realizations of these factors, as well as whether and how they are involved in the individual-specific mechanisms that govern the production of the outcomes of interest, individual treatment effects with respect to one and the same intervention will typically differ between individuals.

This kind of heterogeneity is likely to obtain in many areas traditionally targeted by EBP, e.g. in educational policy, where students may respond differentially to educational initiatives as a function of initial ability; in economic policy where behavioural response to policy interventions may differ significantly between industries, firms and other units of agency; and in public health and development economics, where agents' response to interventions such as distributing free

[2] Standard instances of such variables are so-called treatment effect moderators as well as some kinds of mediating variables.

insecticide-treated bed nets might exhibit substantial heterogeneity as a function of agents' basic needs or epidemiological knowledge.

As these stylized facts indicate, heterogeneity among agents' response to treatment is ubiquitous in several key areas targeted by EBP. Yet, the issue of heterogeneity has only recently attracted attention from EBP proponents (in contrast to evidence-based medicine, see e.g. Oxman and Guyatt 1992 for an early treatment). This is surprising because heterogeneity is also responsible for one of the most basic inferential challenges that EBP faces, i.e. the problem of extrapolating experimental results from study populations to eventual policy targets. Let me expand on some technical background to explain why this is the case.

3.1. Heterogeneity Information from RCTs

More formally, treatment effect heterogeneity is the systematic variation in the sign and/or magnitude of individual treatment effects among agents who are subject to a given intervention. In a potential outcomes framework (Rubin 1974; Holland 1986), given an outcome of interest Y, the individual treatment effect (ITE) for individual i is the difference between her potential outcome $Y_i(1)$ given the treatment and her potential outcome $Y_i(0)$ in the absence of treatment, other things being equal. Since only one of the two values of Y_i can ever be observed, ITEs are in principle unobservable magnitudes (Rubin 1974).

RCTs offer a partial remedy for this inferential dead-end by permitting the estimation of *average treatment effects* (ATEs) instead of ITEs. This is achieved through balancing the net effects of confounding factors as well as interactive covariates by means of random assignment of subjects to experimental and control conditions, and multiple blinding of trial participants, those administering treatment and those recording and interpreting outcomes. Provided that randomization (and blinding etc.) are successful in that the net effects of confounders and interactive covariates (including interactions among them) are approximately balanced between treatment and control groups, an ideal RCT can help obtain, in expectation, an unbiased estimate of the ATE by taking the difference in means of Y for treated and untreated units, or $\widehat{ATE} = \overline{Y_t}(1) - \overline{Y_c}(0)$.

This estimate of the sample ATE, however, does not permit inferences about ITEs. At best, and in the absence of any knowledge about interactive covariates as well as heterogeneity in their individual-specific realizations, the ATE estimate can figure as the expectation of the ITE for an individual randomly drawn from the experimental population. But as soon as there is variation between individuals in their realizations of interactive covariates, and consequently variation in ITEs, inferences about ITEs are

largely precluded and information on heterogeneity cannot be recovered from $\widehat{ATE}$.

This has significant bearing on the *transferability* of trial results, i.e. the extent to which the ATE from an experimental population A can be expected to be replicated in some other population B. Two jointly sufficient conditions for the transferability of trial results (in the particular sense adopted here) to some out-of-sample target are first, that the treatment variable plays the same causal role in the production of the outcome in the target as it does in the experimental population, i.e. that the mechanisms in both populations are sufficiently similar with respect to the causal effect to be extrapolated. The second condition is that the distribution of interactive covariates of the treatment effect is the same in both populations (see e.g. Cartwright and Marcellesi 2015 for such conditions).[3] So the transferability of experimental results to targets hinges not only on sufficient similarity in mechanisms between populations but also on whether there is heterogeneity induced by differences in interactive covariates as well as how these variables are distributed among agents in the populations of interest. This problem has received attention from a variety of econometricians, methodologists, philosophers of science and EBP proponents (e.g. Hotz *et al*. 2005; Duflo *et al*. 2008; Imbens and Wooldridge 2009; Bareinboim and Pearl 2013; Cartwright and Marcellesi 2015).

Treatment effect heterogeneity does not only affect the transferability of trial results. It also creates a second challenge for EBP. The challenge is that in the absence of information on heterogeneity, RCTs are not suitable for informing policy formation processes that are concerned with the *distributive consequences* of policy (see e.g. Manski 2000). More specifically, policymakers may often be interested in knowing not only whether an intervention is effective on average but also in how effective the intervention will be for specific types of agents, how treatment effects are distributed among agents, with respect to which observable pre-treatment characteristics they exhibit heterogeneity, whether heterogeneity obtains in magnitude or also in sign, etc.

Such information is crucial particularly when it is reasonable to suspect that agents may be harmed by an intervention, even though the ATE might be positive. In these scenarios, several pertinent distributive concerns arise, e.g. is it at all permissible to implement a policy that will render some agents worse off? If so, how should we adjudicate between the negative welfare consequences for these agents and the net effectiveness of the intervention? What are the thresholds of proportionality that we should use to decide whether benefits on the part

[3] Necessary conditions might be weaker, see Bareinboim and Pearl (2013).

of some outweigh losses on the part of others? Can the policy be targeted so that it predominantly affects those who will most likely benefit from the intervention? And so forth.

As these considerations suggest, policymakers may be interested in pursuing a variety of different distributive values. Yet, in order to pursue these values effectively, in the sense that the available evidence gives them good reasons to believe that an intervention will in fact promote them, policymakers require information on treatment effect heterogeneity, i.e. whether there is heterogeneity at all and how heterogeneous treatment effects are distributed with respect to agents' observable characteristics. As I have argued above, RCTs cannot provide such information on their own.

This does not mean that EBP methodology is at a complete loss in this regard, however, as one way to address this problem is to perform so-called *subgroup analyses*. While this is feasible as long as researchers have obtained pre-treatment data on potential interactive covariates of the treatment effect, I argue below that performing such analyses, when judged against the background of standard quality-of-evidence ranking schemes circulated in EBP, comes at the expense of sacrificing several key EBP epistemic values. This creates a tradeoff between several epistemic values central to EBP and the pursuit of moral and political values such as equality and priority for the worst-off.

3.2. Subgroup Analysis as a Remedy for Informing about Heterogeneity

Following Duflo *et al*. (2008), subgroup analyses partition experimental populations into subgroups according to observable pre-treatment characteristics such as age, sex, ethnicity, prior education etc. They then typically further partition subgroups into different categories or strata, for instance age groups. Given this stratification, a difference-in-means estimation can be performed on the partitioned data to obtain conditional, subgroup-specific ATEs (CATEs); this helps us tell whether treatment effects differ between subgroups. A somewhat more sophisticated alternative to this are regression-based approaches, where potentially interesting interactive covariates of the treatment effect are modelled as interaction terms with the treatment indicator in a standard regression framework. In doing so, it is possible to obtain information on the significance of interaction effects between observable variables and the treatment indicator, which may be taken to suggest that the variables in question induce heterogeneity in the treatment effect of interest.

Even so, while subgroup analyses provide at least tentative information about heterogeneity, they are also subject to several pertinent methodological concerns. Let me expand on two particularly pressing

concerns and explain how they bear on the achievement of EBP epistemic values.

First, the information that CATE- and regression-based approaches can generate is purely correlational in nature, and hence subject to standard concerns about endogeneity and consequent bias. For instance, consider the case where we estimate a significant positive interaction between microfinance access and prior business ownership, suggesting that agents who previously owned a business will benefit more from microfinance access. Such a finding does not permit the straightforward conclusion that prior business ownership is a *causally relevant* interactive covariate of the treatment effect. This is because the significance of the estimate may be attributable to common-causes, e.g. because business ownership is strongly correlated with business education, and it is business education that is causally relevant for inducing different behavioural responses to microfinance access, but prior business ownership in the absence of (or conditional on) business education may not contribute at all to microfinance treatment effects.[4] In this case, if business education is not included in the regression, our estimates of individual-level heterogeneity with respect to prior business ownership will be upwards biased.

Randomization at the treatment stage does not alleviate this problem because although it ensures, in expectation, that the net effects of prior business ownership and business education on the outcome are distributed equally between treated and untreated units, it leaves the covariance *between* the two variables untouched. So if we run a regression with an interaction term including only prior business ownership and find that there is a significant interaction with the treatment indicator, this result can be misleading. This is because the variable that truly induces the subgroup differences, business education, will be captured by the error term, and since it is still correlated with prior business ownership, this yields a biased estimate of the interaction between treatment and prior business ownership.

More generally, parameter estimates for interactive treatment effect covariates will invariably remain subject to such concerns unless researchers are prepared to entertain the relatively strong assumption that

[4] For instance, prior business ownership in the absence of business education can be exhibited by agents who have previously pursued unprofitable business plans and may continue to do so in the future. Thus the unbiased parameter estimate for business ownership is likely to be substantially smaller than the estimate for business education. To permit unbiased estimation of interaction terms, one would at least need to induce additional exogenous variation in the covariates of interest. But this would require significantly different trials designs with multiple, parallel interventions on treatment as well as interactive covariates (see e.g. Imai *et al*. 2013). While such designs are in principle feasible, they also raise issues with precision and statistical power.

the subgroup variables of interest are uncorrelated with the error term of the regression (cf. Pearl 2014), i.e. there may be no common causes of the putative subgroup variable of interest and the outcome that are captured by the error term and that could induce the apparent interaction of the subgroup variable with the treatment indicator. However, it is precisely such questionable identification assumptions, which are necessary for unbiased identification in any standard regression context, that EBP proponents are keen to avoid. Indeed, randomization is expressly emphasized in the methodological literature as the key strategy to help avoid making such assumptions.

This means that using subgroup analyses to obtain *unbiased estimates* of heterogeneous treatment effects and straightforward *causal conclusions* about the role of interactive covariates that induce them is typically precluded, threatening at least two EBP epistemic values at once.

A second worry about subgroup analyses is with regard to the *precision* of effect estimates, including concerns about *statistical power*. In short, the more subgroups one specifies, the higher the probability of obtaining spurious results. For typical significance levels at $P < 0.05$ even a moderate number of subgroups, strata partitions and corresponding hypothesis tests will render the occurrence of spurious results exceedingly likely. At the very least, suitable statistical corrections are in order to remedy the consequences of multiple testing for the prevalence of false positives. Yet, while recommended by some EBP proponents (e.g. Duflo *et al*. 2008: 65), this is rarely carried out in practice (Fink *et al*. 2014: 47). In short, these concerns about hypothesis testing pose a threat to the envisioned *methodological rigour* mandated in standard EBP methodological guidelines.

Moreover, to alleviate concerns about insufficient statistical power, sample sizes may need to be expanded for subgroup analyses to be informative at all. For instance, in order to detect a heterogeneity signal of the same magnitude as the sample ATE and with the same precision as the sample ATE estimate, a difference-in-means estimation on just one subgroup partitioned into two strata requires a fourfold expansion of the original sample size (Varadhan and Seeger 2013: 38). Yet, subgroup-specific effects may be smaller than ATEs, particularly in relatively homogeneous trial populations.[5] So adequately powered subgroup analyses may frequently require much greater expansions of sample sizes to permit detection of heterogeneous effects. In short, this suggests that the epistemic value of *precision* is threatened for the investigation of subgroup-specific effects.

[5] This does not mean that heterogeneity in eventual target populations, as well as heterogeneity between experimental and target populations is similarly mild, however, so it is still important to learn about heterogeneous effects.

These and other, related concerns limit the extent to which subgroup analyses can inform about treatment effect heterogeneity in a way that lives up to the epistemic standards imposed by EBP methodological guidelines. Standard methodological recommendations (e.g. Varadhan and Seeger 2013) suggest that subgroup findings should at most be considered *exploratory* in the sense that they may prompt additional investigations such as novel trials on subgroups of interest, but are insufficient to warrant definitive conclusions about heterogeneity by themselves. However, while conducting novel RCTs on subgroups appears to be a viable strategy for addressing some of the above concerns, this requires prior identification of the relevant subgroups. Unfortunately, we are rarely in the epistemically fortunate position to know which individuals are most likely to incur welfare losses in advance, since that depends on knowing what the causally relevant interactive covariates of the treatment effect are, how they affect the outcomes of interest as well as which agents exhibit beneficial or harmful realizations of these variables. So information on heterogeneity, possibly obtained from subgroup analyses, is still required even if we are willing to conduct subsequent RCTs on particular subgroups to obtain unbiased estimates of subgroup-specific effects.

The extant EBP literature has only recently started to address treatment effect heterogeneity issues. Yet, even though there are several recent social policy and development studies that perform at least tentative and exploratory heterogeneity analyses, they frequently fail to address one or more of the concerns outlined above (see e.g. Fink *et al.* 2014) or tend to focus on heterogeneity that obtains between estimates obtained in *different* trials, which is a related but conceptually distinct issue from the *within-trial* and *between-subject* heterogeneity that I consider here. Specifically, between-trial heterogeneity may not only occur as a result of individual-level differences in treatment effects, but also as a result of trial-specific differences such as differences in the interventions being tested or the quality of treatment implementation, as well as differences in the methods used to estimate their effects. Investigations of such between-trial heterogeneity are mostly focused on determining whether heterogeneity is random or systematic, attributing heterogeneity to differences between trials, and deciding whether trial results, despite such differences, may still be aggregated in systematic reviews of effectiveness evidence. Such studies are hence not typically concerned with exploring individual-level heterogeneity in treatment effects and even less so with the distribution of treatment effects within populations.

Let me expand on how these epistemic challenges for informing about heterogeneity create a trade-off between epistemic and moral values in EBP and how this trade-off challenges both value-freedom and neutrality in EBP.

4. A TRADE-OFF BETWEEN EPISTEMIC AND NON-EPISTEMIC VALUES

The trade-off between epistemic and non-epistemic values that I want to highlight is a result of the differential usefulness of EBP research outcomes for the pursuit of different kinds of moral values, i.e. values that do or do not take into consideration the distribution of harms and benefits induced by policy interventions.

Standard EBP methods such as RCTs are in general capable of estimating ATEs.[6] These quantities are sufficient for the pursuit of values that are concerned with increasing or maximizing aggregate or average welfare, such as the kinds of broadly utilitarian values pursued by policymakers who focus on the net (cost-) effectiveness of policy interventions. Standard EBP evidence is suitable for the pursuit of such values because the distribution of individual-specific contributions to aggregate effects is not a primary concern when aiming to increase aggregate or average welfare, so information on heterogeneity is not necessary for the pursuit of these values.[7]

However, information on heterogeneity is necessary for the pursuit of any moral and political value that is sensitive to *how* aggregate outcomes are realized. For instance, the pursuit of broadly egalitarian or prioritarian values requires at least information on the pre-intervention distribution of the outcome variable among agents as well as information on the changes to this distribution brought about by the intervention at issue.[8] Similarly, pursuing a strict paretian welfare criterion, a precautionary principle, or any other value that places particular emphasis on not harming agents, will require one to obtain information on whether any agents are made worse off by some intervention. As I have argued above, such information cannot be provided by RCTs alone. At the very least, subgroup analyses need to be carried out in order to permit at least tentative conclusions about heterogeneity. Unfortunately, standard EBP methodological recommendations tend to explicitly discourage subgroup analyses.

For instance, JPAL, one of the leading institutions in development programme evaluation, cautions against the use of subgroup analyses. In the methodology section on their website, JPAL warns specifically

[6] This extends to quasi-experimental approaches such as matching, instrumental variables and regression discontinuity identification strategies as well. While some of these approaches can only (without strong assumptions) identify *local* average treatment effects (LATEs), the concerns outlined here apply to these approaches as well since the causal quantities at issue are still *average* quantities.

[7] It might still be helpful, since welfare *maximization* may be easier to accomplish when we have information that helps pick out those individuals who will likely benefit most from some intervention.

[8] See Atkinson (2011) for similar concerns about the limited ability of representative agent models to inform about distributive consequences of interventions.

about the occurrence of spurious results, i.e. that when testing multiple subgroup hypotheses 'it is likely that the [subgroup] difference[s] [are] due simply to random chance – not our program' (JPAL 2017). While it is important to emphasize the problems associated with multiple hypothesis testing, JPAL's methodology section does not offer advice for how to correct for multiple testing, which statistical methods for estimating heterogeneous treatment effects are preferable, or how the credibility of subgroup analyses should be assessed when relevant precautions are taken. Absent such guidance, this suggests that ATEs constitute distinctively superior evidence for policy design purposes.

In a similar vein, the What Works Clearinghouse (WWC) Procedures and Standards Handbook (v.3.0) for conducting systematic reviews expresses a clear preference for aggregate quantities, i.e. ATEs, over subgroup-specific results:

> When a study presents findings separately for several groups … without presenting an aggregate result, the WWC will query authors to see if they conducted an analysis on the full sample ... . If the WWC is unable to obtain aggregate results from the author, the WWC averages across subgroups within a study to use as the primary finding and presents the subgroup results as supplemental tables. (What Works Clearinghouse 2014: 17)

Moreover, for expedited reviews, the WWC exercises 'discretion to focus each study review on eligible findings only from the full sample (rather than on subgroups)' (What Works Clearinghouse 2017: 18), suggesting that a trade-off between the expediency of a systematic review and the informativeness of the results about the distribution of treatment effects is settled in favour of expediency.[9]

Finally, the WWC does not consider subgroup analysis evidence relevant for overall assessments of programme effects. Specifically, '[f]or WWC intervention reports, the average measure factors into the intervention rating, but the separate subgroup results do not' (What Works Clearinghouse 2014: 28), meaning that overall assessments of comparative effectiveness of interventions simply disregard the distributive consequences of these interventions.

Similarly, according to the CONSORT 2010 statement, subgroup analyses are *ancillary analyses*, which means that they are not considered to be part of the main results of an effectiveness analysis; thus again placing recognizably more emphasis on ATE results. In a more general assessment, CONSORT also cautions that

> because of the high risk for spurious findings, subgroup analyses are often discouraged. Post hoc subgroup comparisons (analyses done after looking

[9] See Elliott and McKaughan (2014) for a related case on the relationship between epistemic values and non-epistemic values such as expediency.

> at the data) are especially likely not to be confirmed by further studies. Such analyses do not have great credibility. (Moher *et al*. 2010: 14)

Another pertinent example comes from the Cochrane Collaboration Handbook,[10] which cautions that subgroup analyses may be misleading since they are observational in nature and 'suffer the same limitations of any observational investigation, including possible bias through confounding' (Higgins and Green 2011: ch. 9.6.6). This makes clear that subgroup analyses are generally considered as being of the same quality and credibility as observational studies, which typically rank lower in quality-of-evidence rankings than the ATEs reported in the same studies that such subgroup analyses may be part of. Again, the trade-off between informativeness regarding distributive consequences of interventions and the potential bias involved in subgroup analyses that could produce such information is settled in favour of *unbiasedness*.

Another important set of recommendations comes from the GRADE guidelines for systematic reviews, which consider treatment effect heterogeneity and subgroup analysis under the general heading of *factors that reduce the quality of evidence*. GRADE considers treatment effect heterogeneity under the label of *inconsistency*, where evidence is considered inconsistent if multiple studies on the same or similar interventions produce different point estimates of treatment effects, thus indicating potential heterogeneity in treatment effects.

There are two notable recommendations in the GRADE guidelines. The first is that authors of systematic reviews should 'combine results [in a systematic review] only if … it is plausible that the underlying magnitude of treatment effect is similar' (Guyatt *et al*. 2011: 1295). So different point estimates of treatment effects should only be combined if there is little or no heterogeneity in the treatment effects across studies. This makes it unlikely that systematic reviews will be able to provide comprehensive accounts of heterogeneous treatment effects since heterogeneity itself is taken as reason not to consider the evidence-base as amenable to a single systematic review.

Moreover, according to GRADE, heterogeneity in treatment effects, if unexplained, should be taken as reason to discount the quality of a body of evidence (Guyatt *et al*. 2011: 1295). While GRADE suggests that authors should try to explain apparent heterogeneity by means of subgroup analyses, GRADE also maintains that ' most putative subgroup effects ultimately prove spurious' (Guyatt *et al*. 2011: 1297).

[10] While the Cochrane Collaboration focuses on evidence-based medicine, it is still widely considered to provide useful guidelines for effectiveness evaluation of interventions more generally. For instance, the Campbell Collaboration guidelines, which focus on Evidence-Based Policy, make extensive references to the recommendations offered in the Cochrane Handbook.

This seems odd. Surely, there can be cases where apparent heterogeneity is spurious or arises due to errors, including cases where it is difficult to tell that this is so. In these cases it would seem sensible to discount the quality of a body of evidence. At the same time, there will also arguably be many cases where there is genuine heterogeneity in the effect of interest, including cases where important heterogeneity might remain unexplained, despite our best efforts to explain it. What is striking here is that, at least in these latter cases, the GRADE recommendations would imply that the very nature of the phenomenon under scrutiny should be taken as a reason to discount the quality of the evidence pertaining to the phenomenon. It hence seems that the GRADE recommendations are only sensible if one believes that unexplained heterogeneity is significantly more likely to be spurious or a result of error, rather than genuine but unexplained despite our best efforts. It remains unclear, however, what the argument for this presupposition is, as well as what underlying standard GRADE envisions when it comes to classifying heterogeneity as unexplained. In the absence of further clarifications on these issues, it seems that the GRADE recommendations establish a clear, but unsubstantiated preference to discount evidence indicating potentially important heterogeneity, as well as the subgroup evidence that could help elucidate such heterogeneity.

The above methodological recommendations clearly signal that subgroup analyses are generally considered to enjoy significantly less credibility than ATE results, are ranked lower in terms of quality of evidence, and are explicitly bracketed from overall effectiveness evaluations in systematic reviews. What is more, most methodological recommendations focus on highlighting potential problems with subgroup analyses, but remain entirely silent on the importance of subgroup evidence for welfare analysis.[11]

In addition, alternative methods for detecting and attributing heterogeneity such as machine learning methods that promise efficient and unbiased detection of heterogeneity from large-N observational data (e.g. Athey and Imbens 2016, 2017) are rarely acknowledged or mentioned in standard methodological guidelines. Even if they were, these methods would not be straightforwardly compatible with the standard experimental designs involving relatively small samples that

[11] The Campbell Collaboration's Conduct Standards guidebook is an exception at least insofar as it suggests that it is highly desirable, although not mandatory, for reviews to 'include explicit descriptions of the effects of the interventions not only on the whole population but also describe their effects upon specific population subgroups and/or their ability to reduce inequalities and to promote their use to the community' (Campbell Collaboration 2016: 3). At the same time, this suggestion is still at odds with the Campbell Collaboration's extensive references to the Cochrane Collaboration Handbook, which explicitly cautions against the use of subgroup analysis.

EBP researchers typically employ, nor would they rank highly on the quality-of-evidence rankings that are circulated in the methodological literature as they would, again, be considered observational studies and hence would be considered to provide evidence of distinctively less quality and credibility than RCTs.

This suggests that the trade-off between the informativeness of evidence concerning distributive issues and the epistemic values involved in assessing the quality of evidence is presently settled, at the level of several widely disseminated bodies of methodological recommendations, in favour of the usual standards and the epistemic values that they are supposed to promote. Following these recommendations hence privileges the production and use of ATE evidence as the kind of evidence that may ultimately form a credible basis for policy design.

This licenses two conclusions. First, EBP methodology presently favours the production and use of ATE evidence that is useful for the pursuit of values that focus on increasing average or aggregate welfare. Second, EBP methodology presently fails to promote or even discourages the production of evidence that is necessary for the pursuit of many values that put emphasis on the distribution of treatment effects. As a consequence, standard EBP methodology renders the pursuit of many values on grounds of EBP evidence relatively more difficult or even outright infeasible.

To be clear, this is not to say that ATE evidence cannot be useful *at all* for the pursuit of values that focus on distributive issues. For instance, a large negative ATE might strongly suggest that a certain intervention should not be implemented; and this conclusion might be action-guiding irrespective of the particular values one is interested in pursuing. Similarly, if one estimates a particularly large and positive ATE, this may, in some cases, give us reasons to think that an intervention is at least somewhat unlikely to make any individuals worse off.

At the same time, even in such favourable cases, the *differential* usefulness of ATE evidence for the pursuit of different kinds of values still persists. Specifically, if one cares only about average effects, then the only important quantity is the ATE, which is usually well-identified in RCTs. On the other hand, putting particular emphasis on whether an intervention has adverse effects on any individuals will typically require focusing on information other than the ATE, e.g. results obtained from subgroup analyses or pre-post intervention comparisons of outcome distributions. Such analyses can generally not yield unbiased estimates of individual treatment effects, so while one can perform such additional analyses, they do generally not enjoy the same credibility as the main ATEs reported in effectiveness evaluations. This still implies that standard EBP evidence makes the pursuit of values that are sensitive to the

distribution of treatment effects *relatively* more difficult, although perhaps not always infeasible.

This can have various undesirable consequences. To illustrate, consider the case of two policymakers A and B. Suppose that A is concerned with increasing average or aggregate outcomes, whereas B pursues prioritarian values, i.e. she cares specifically about whether an intervention promotes the outcomes of the pre-intervention worst-off individuals. If the available evidence on policy effectiveness reports mostly the ATEs of interventions, then B will be in a worse position than A to justify her calls for policy action. This is because the information that she needs to justify these calls is either not available at all, or, if available, e.g. in the form of subgroup analyses, is considered to be of less quality than the ATE evidence that A can invoke to justify her calls for policy action. Facing standard quality-of-evidence rankings, B will hence find it recognizably more difficult to justify her calls for policy action, given one and the same evidence-base. For instance, political opponents may find it easier to question the credibility of the evidence that B invokes, as standard quality-of-evidence guidelines discount the credibility of subgroup results, thus making it more difficult for B to resist such scrutiny. Moreover, scrutiny of subgroup evidence on epistemic grounds may also allow non-epistemic motivations to creep back into evidence-based policymaking. For instance, when an opponent challenges subgroup-evidence on purportedly epistemic grounds she might in fact do so because she considers the policy that is being justified by appeal to such evidence undesirable for moral and political reasons (see e.g. Barnes and Parkhurst 2014; Parkhurst and Abeysinghe 2016 for similar concerns). Finally, this situation may also incentivize B to shift the values she will ultimately promote to those for which high-quality evidence is available, e.g. by putting more emphasis on average effectiveness rather than effectiveness that is construed in accordance with her prioritarian values.

This situation suggests that there is presently a trade-off between several epistemic values central to EBP and the moral and political values that policymakers are in a position to pursue effectively on grounds of EBP evidence. More specifically, whenever the pursuit of moral and political values requires information about the distributive consequences of policies, standard EBP evidence fails to provide the required information. Conversely, whenever evidence of the kind required to inform about distributive consequences of policies is produced or used, this typically involves sacrificing at least some EBP epistemic values. More specifically, as argued above, when methods such as subgroup analyses are used to generate information on treatment effect heterogeneity, this may come at the expense of sacrificing the *unbiasedness* and *precision* of effect estimates, the *methodological rigour* in obtaining such estimates,

as well as the *ability to obtain causal conclusions* on the basis of such estimates. Insisting on these values, on the other hand, comes at the expense of sacrificing the informativeness of EBP research outputs about the distributive consequences of policy interventions.[12]

It is important to be clear that this trade-off only obtains if there is a strong, and perhaps unique relationship of fit between certain methods such as RCTs and the epistemic values that I take to be underlying EBP methodology. If that were not the case, then there could be other methods that manage to promote the same set of epistemic values; potentially including methods that are more informative about distributive issues.

I am open to this possibility, as it seems that the trade-off outlined here is, at bottom, not a necessary one, but one that holds contingently upon specific features of EBP methodology. More specifically, it seems possible to hold onto at least some EBP epistemic values while using methods that are more informative about distributive consequences. For instance, one can be less or more *rigorous* when conducting subgroup analyses, e.g. by taking adequate precautions such as pre-specifying subgroup hypotheses to ameliorate concerns about data mining. Moreover, subgroup analysis can also yield more *precise* estimates of subgroup-specific effects if one has sufficiently large samples. So there is no necessary compromise of some or even all EBP epistemic values when using such methods.

At the same time, the particular ways in which different methods do in fact strike a balance between promoting certain epistemic values and producing informative evidence on the distributive consequences of interventions is only part of what constitutes the trade-off I am interested in. Another, and arguably more important part is the way in which the relationship of fit between epistemic values and methods is conceived in standard EBP methodological recommendations. In virtue of lexically prioritizing certain methods and certain kinds of evidence, as is common in widely disseminated quality-of-evidence ranking schemes, standard EBP methodological recommendations suggest that there is a strong, and perhaps unique relationship of fit between methods such as RCTs and epistemic values such as unbiasedness, precision etc. So what I am arguing is not that there is a necessary trade-off, but that this trade-off obtains primarily relative to how the ability of different methods to

[12] This point may appear similar to Helen Longino's, who argues that several traditional epistemic values are not purely epistemic and 'that their use in certain contexts of scientific judgment imports significant socio-political values into those contexts' (Longino 1996: 54). In contrast, my point should appeal even to those who insist on the purely epistemic character of values such as unbiasedness, precision, and the ability to obtain causal conclusions. Specifically, I do not argue that these values fail to be purely epistemic. Instead, even if we grant that they are purely epistemic, their pursuit may still have important ramifications for the extent to which the pursuit of other, moral values is facilitated or inhibited.

promote the achievement of key EBP epistemic values is conceived in EBP methodological recommendations.

With this overview of the trade-off between epistemic and moral values in place, let me expand on what it implies for value-freedom and neutrality in EBP respectively.

4.1. Value-freedom

First, if the value-free ideal underlying EBP is understood as saying that non-epistemic values are generally not and should not be involved in shaping the conduct and outcomes of EBP research internally, e.g. by influencing the collection of data as well as the choice between different methods, estimators, model specifications and so forth, then the desirability of this ideal is challenged. Without suitable changes to EBP methodology, pursuing values that depend on the distribution of treatment effects on grounds of EBP evidence is presently inhibited. If this situation should be remedied, then this requires changes to EBP methodology that enable and facilitate the production and use of evidence on treatment effect heterogeneity. However, and this is the crucial point, these changes would be effected by *moral values*, since it is the desired ability to effectively pursue moral values on grounds of EBP evidence that motivates the requisite changes to methodology. To the extent that such changes to methodology are justifiable and justified, this suggests that value-freedom in EBP is not a desirable ideal, even at key methodological stages such as issuing widely disseminated recommendations for method choice, data collection, model specification, estimation and interpretation.

It is important to emphasize that this may be a transient state of affairs only, because as soon as requisite changes to standard methodological recommendations are implemented, and the evidence that is produced in accordance with revised recommendations becomes more informative about distributive consequences of policy interventions, then this may obviate further changes to methodology that are driven by non-epistemic, moral values. So value freedom in EBP may only be *transiently* undesirable. This helps to push back against the concern that the changes to EBP methodology suggested above are just the first step in permitting arbitrary influences of moral values on the internal stages of EBP research.

To further push back on such concerns, it is important to emphasize that not all changes to EBP methodology will promote the goal of increasing the extent to which EBP evidence is informative for the pursuit of widely shared moral and political values such as equality and priority for the worst off. This goal is sufficiently specific to rule out changes to EBP methodology that are motivated by other, potentially idiosyncratic values, e.g. researchers' personal convictions about the desirability of

policy interventions, or profit-maximization motives by trial sponsors who prefer methods that are more likely to produce effect estimates that are in accordance with their financial interests. Such value-influences would not be compatible with the kinds of changes to EBP methodology that facilitate the pursuit of a broader set of moral and political values on grounds of EBP evidence. So value-freedom in EBP may still remain an important ideal, but there are reasons to think that it is presently undesirable, as the pursuit of an important class of values on grounds of EBP evidence is inhibited until EBP methodological tenets undergo suitable revisions.

The role for values outlined here hence differs, and extends considerably beyond extant contributions such as Heather Douglas's (2009), who argues that moral values play important and ineliminable roles in handling the *uncertainties* that are involved in using evidence for policy purposes; and that such values may play legitimate *indirect* roles in the *assessment* and *use* of uncertain evidence. The arguments provided here suggest that there are cases where the role of non-epistemic moral values may be even more extensive than previously considered as they may also, legitimately, play *direct* roles in governing the *production* of evidence, e.g. when a particular moral value figures as a reason in itself (cf. Douglas 2009: 96) to recommend a particular method over another, or to use a method in a certain way that helps promote the pursuit of the moral value in question, even if this proceeds on pain of sacrificing other, epistemic values.

Two important qualifications need to be added here. First, it is clear that there can be cases in which policymakers can offer precise characterizations of the questions they are interested in answering, and with a view towards the particular values that they are interested in pursuing. In these cases, when the questions to be addressed are fixed, it seems that even if values were involved in shaping the questions that are pursued by EBP researchers, they would not meddle with the outputs of that research. This seems a clear-cut case where values play only an external role.[13]

My concerns apply to a different type of case, however: cases where evidence is used 'off the shelf'. This relates to an important goal of EBP, i.e. to build libraries of evidence that policymakers and practitioners can consult when addressing certain types of policy problems, or implementing certain kinds of interventions. In these cases, evidence is produced before the aims of the particular use-case are determined, including any values that are to be pursued. Here, it seems that non-epistemic values may play direct, and legitimate roles at internal stages. They can act as reasons in themselves, explicitly mentioned

[13] I thank an anonymous referee of this journal for raising this important concern.

in methodological recommendations, and on par with other, epistemic values, for choosing certain methods over others, even if this comes at an epistemic price, e.g. less precise estimates, or increases in the risk of bias.

Of course, this does not mean that any sort of wishful thinking is going on, as for instance when certain methods are chosen on the grounds that they are more likely to produce particular kinds of answers. But that is precisely how my case departs from Douglas's concern about wishful thinking cases, i.e. cases where values systematically and illegitimately meddle with our research results. In contrast, in the present case it seems that even though values play a direct role at internal stages, such as method choice and model specification, this is not necessarily illegitimate. The reason is that the choice of method may have important non-epistemic ramifications concerning the kinds of policies that are likely to be implemented on grounds of EBP evidence, and these ramifications may, legitimately, be anticipated when governing the production of evidence.

This points to a second important qualification, which is that I am not suggesting that moral values need to act as reasons in themselves at internal stages of particular studies. My concerns are rather with the methodological recommendations that are circulated in EBP, and that shape the conduct of individual studies. Hence, we might say that value-freedom in EBP is transiently undesirable, not at the level of individual studies, but rather at the methodological level of issuing recommendations for the production and use of different kinds of evidence. In analogy to study-level internal roles for values in determining what method to use, which model specification to use, how to collect data, and how to interpret results, I am concerned with internal roles for values at the level of general methodological recommendations pertaining to these issues. Hence, my claim is that value freedom is, at least transiently, undesirable at the level of widely disseminated methodological recommendations, although perhaps not at the level of individual studies.

Let me expand on related concerns about value-neutrality in EBP.

4.2. Value-neutrality

Recall that for the purposes of this paper I understand value-neutrality in EBP as the idea that EBP research outputs should not be premised on substantive moral value judgements about the desirability of social outcomes or the interventions that bring about these outcomes. So in disseminating the results of policy evaluations EBP researchers may issue at most *conditionally* normative recommendations, i.e. recommendations that are conditional on substantive value judgements but do not endorse

these judgements as such. For instance, when a policy is considered most (cost-) effective because it best promotes an outcome such as household endowment of the rural poor, nutritional health in children etc. it may be explicitly recommended on grounds of its effectiveness. However, whether or not the outcomes that the intervention best promotes are in fact morally desirable should be settled independently, and EBP policy recommendations should remain neutral with respect to such questions. This issue is for policymakers and other suitably legitimized agents to settle.

As the previous discussion suggests, this type of value-neutrality is undermined by how the trade-off between the informativeness of evidence and central EBP epistemic values is settled in practice. More specifically, adherence to widely disseminated methodological recommendations and the epistemic values underlying them means that inferences about policy effectiveness remain limited to average effectiveness assessments and hence do not encompass information about the distributive consequences of interventions. So the standard way of operationalizing what it means for a programme to be *effective* brackets concerns about distributive issues (see Biller-Andorno *et al*. 2002 for similar concerns about evidence-based medicine). As it stands, an *effective* programme is considered a good programme to the extent that the outcome of interest tracks a relevant moral or societal good. However, even if this good is uncontroversial in itself, effectiveness as standardly construed in EBP still only means effectiveness on average, not *some* effectiveness for everyone, or *sufficient* effectiveness for the worst-off, or *equal* effectiveness for all policy subjects. It is clear that effectiveness in one sense does not imply effectiveness in others, so policies that are effective on average can have distributive consequences that are undesirable relative to a wide variety of values, and hence would not be considered effective relative to these values.

To achieve value-neutrality in the envisioned sense it is hence not enough to ensure that the desirability of the outcomes of interest *as such* has been settled, or can be settled independently. It is also necessary to maintain neutrality with respect to the *ways in which* these outcomes may be realized. A given change in aggregate outcomes can usually be achieved in various ways, each of which may have dramatically different distributive consequences for target populations, and some of these may be intrinsically more or less desirable than others. If policymakers who wish to use EBP evidence care about differences between agents, and about absolute and relative changes in outcome distributions, then *effectiveness* as standardly construed in EBP is not informative about the moral permissibility or desirability of the policies under scrutiny and might be misleading about what *effective* programmes are ultimately

able to achieve, relative to the specific moral and political values that policymakers pursue.

So, at least presently, it seems that the dissemination of EBP research is premised on the implicit presupposition that the relevant magnitude for deciding which policy to implement, is effectiveness in terms of average effects. This fails to be value-neutral in the envisioned sense because it presupposes that *average* effectiveness is the magnitude of interest to policymakers, rather than delegating the question of whether it is to policymakers and other suitably legitimized agents to settle. In a nutshell, in order to maintain a traditional ideal of value-neutrality, additional value presuppositions such as the above must at least be made explicit for EBP policy recommendations to remain value-neutral in the envisioned sense.

Let me close with some general remarks on the role of non-epistemic values in EBP going forward.

4.3. EBP and Values – Where Next?

The previous discussion raises important questions about what role non-epistemic moral and political values *should* play in EBP, and consequently how potential tradeoffs between epistemic and non-epistemic values should be settled at the level of widely circulated methodological recommendations. Should EBP proponents give up on their commitment to central EBP epistemic values, and if so, which epistemic values? Or should they bite the bullet and concede that standard EBP evidence fails to be informative for the pursuit of a wide range of moral and political values?

It seems unclear whether there is a single, univocal answer to how this trade-off should be settled. It seems plausible to think that the precise nature of this trade-off will depend on concrete contextual details pertaining to the kinds of questions that users of evidence seek to address, the nature of the policy settings that these questions pertain to, and the nature of the methods that are available for addressing these questions; specifically their relative ability to provide certain kinds of information and the extent to which they promote certain kinds of epistemic values. These aspects can vary importantly between cases.

What I want to offer here is hence not a set of definitive answers, but rather some general suggestions for moving forward. Some obvious suggestions include that the problems outlined above should be explicitly recognized in EBP methodological guidelines, including an explicit discussion of the differential usefulness of evidence for the pursuit of different purposes and values, as well as recognition of the fact that evidence concerning heterogeneous treatment effects is crucially important for the pursuit of a wide range of values and purposes. With

the importance of such evidence more clearly emphasized, it would also seem desirable if more efforts were exerted in the way of offering up to date advice on recently proposed methods such as machine learning approaches for detecting heterogeneity (e.g. Athey and Imbens 2016, 2017) including recommendations for how these methods should be used, and how the evidence supplied by them can be integrated to yield more comprehensive assessments of the distributive consequences of policy interventions. Moreover, it seems sensible to suggest that EBP methodological recommendations should offer explicit guidelines for how authors of primary EBP research and meta-analyses should comment on potential limitations of their research. For instance, one may recommend that authors expand more explicitly on the deliberations underlying their choice of specific methods, including, if applicable, a commentary on why they chose not to use certain methods and how and why they may have chosen to sacrifice informativeness of the evidence in the pursuit of adherence to certain epistemic values. This is similar to what Heather Douglas (2009) and many others (e.g. Elliott 2017) have been calling for, i.e. more transparency on the part of researchers about choices that may invariably involve non-epistemic considerations.

My second, and more general suggestion is that a general framework is needed to help facilitate deliberation about the role of values in EBP, i.e. a framework that can help structure and elucidate different ways in which epistemic and non-epistemic values may play a role in EBP methodological recommendations going forward.

In standard EBP methodological recommendations, the production of different kinds of evidence is presently recommended on epistemic grounds, i.e. evidence-ranking schemes rank different types of evidence according to their ability to promote standard EBP epistemic values. As the previous discussion makes clear, however, even evidence that ticks all the boxes on the epistemic desiderata underlying such rankings will neither necessarily, nor typically, be useful for the pursuit of all important purposes. This suggests that there is not only an epistemic dimension to recommending the production and use of certain kinds of evidence, but also a non-epistemic dimension concerning what kinds of evidence are useful for the pursuit of different values and purposes.

Recognizing the importance of both dimensions suggests that it may be useful to adopt a framework that integrates both epistemic and non-epistemic considerations relevant for governing the production and use of evidence for EBP. This is not an entirely new idea, of course. There have been several proposals in the philosophy of science literature to apply so-called *coupled epistemic-ethical frameworks* (e.g. Tuana 2013) to various policy-relevant activities in the special sciences, e.g. in climate science (Tuana *et al*. 2012) and evidence-based medicine (e.g.

Katikireddi and Valles 2015). While I am not suggesting that these frameworks can be straightforwardly applied to the present case, it seems that integrating epistemic and ethical concerns might be useful for investigating, structuring and eventually settling the above tensions between epistemic and non-epistemic values in EBP.

There are various ways in which this can be fleshed out in more detail. For instance, one way could be to say that in addition to ranking evidence according to its quality on epistemic grounds, there should be complementary attempts to rank different kinds of evidence according to their ability to individually, or jointly with other kinds of evidence, provide comprehensive accounts of policy effects, including details on adverse effects on subgroups as well as changes in the outcome distributions brought about by policy interventions.

These two dimensions together may allow us to systematically explore the particular balance that different kinds of evidence, and different combinations of such evidence, strike concerning the extent to which they promote epistemic and non-epistemic desiderata respectively. Some methods might do well on epistemic desiderata but may be extremely limited in their scope of sensible application; other methods might strike a more even balance. Analysing such trade-offs systematically, across a range of different methods and different types of policy scenarios, may help us get a better grasp of the joint epistemic and non-epistemic consequences of issuing preferences for certain kinds of methods over others. Finally, a joint epistemic-ethical analysis may also help us to distinguish between different kinds of trade-offs that vary in severity as well as cases where trade-offs may perhaps not obtain at all.

Naturally, considering both epistemic and non-epistemic dimensions in recommending the production of particular kinds of evidence will raise the important question of how these dimensions should be weighed against each other. Any particular weighting of these dimensions will reflect some way in which a trade-off between epistemic and non-epistemic, moral values is settled. As I have suggested above, my aim is not to recommend specific weightings, but merely to highlight that it seems important to develop a framework that facilitates deliberation about such weightings by making choices pertaining to them explicit. So similar to other calls in the values in science literature to make the role of non-epistemic, moral values as explicit as possible, my suggestion here is that adopting an epistemic-ethical approach can mark an important first step in facilitating joint deliberation among methodologists, researchers, users of EBP evidence and relevant stakeholders about how trade-offs between quality of evidence on epistemic grounds and the usefulness of evidence for the pursuit of different values should be settled.

5. CONCLUSIONS

I have argued that there is a trade-off between several key EBP epistemic values and non-epistemic, moral and political values that are sensitive to distributive consequences of policies, e.g. equality and priority for the worst-off. This trade-off obtains because the outputs afforded by standard EBP methods are differentially useful for the pursuit of different moral and political values. I have argued that this trade-off challenges, at least transiently, ideals of value-freedom and neutrality in EBP. This may be taken as a starting point to reconsider, in an epistemic-ethical framework, some of the standard epistemic value presuppositions entertained in EBP. Doing so, I hope, can help refine EBP methodological recommendations in ways that enable and facilitate the production of evidence that is useful for pursuing a wider range of important moral and political values.

ACKNOWLEDGEMENTS

I would like to thank two anonymous referees of *Economics and Philosophy*, as well as Wendy Parker, Julian Reiss, Nancy Cartwright and members of the CHESS and K4U research groups at Durham University for their many helpful comments and suggestions on earlier drafts of this paper. I would also like to thank the audiences at PSA2016, the Science, Values and Democracy workshop at Tilburg University, VMST6 at UT Dallas, the YSI Young Scholars Initiative Workshop at INEM 2017 and the DGPhil 2017 conference for raising several important points that helped refine my arguments. Finally, I would like to thank my students in 'Evidence-Based Policy' and 'Values in Science' at the University of Bayreuth for many stimulating discussions of the ideas presented here. My work on this paper was financially supported by an AHRC Northern Bridge Doctoral Studentship (grant number: AH/L503927/1) and a Durham Doctoral Studentship, for which I am very grateful.

REFERENCES

Atkinson, A. B. 2011. The restoration of welfare economics. *American Economic Review* 101: 157–161.

Athey, S. and G. W. Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *PNAS* 113: 7353–7360.

Athey, S. and G. W. Imbens. 2017. The state of applied econometrics: causality and policy evaluation. *Journal of Economic Perspectives* 31(2): 3–32.

Banerjee, A., E. Breza, E. Duflo and C. Kinnan. 2017. Do credit constraints limit entrepreneurship? Heterogeneity in the returns to microfinance. Buffett Institute Global Poverty Research Lab Working Paper No. 17–104. <https://ssrn.com/abstract=3126359>

Bareinboim, E. and J. Pearl. 2013. A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference* 1: 107–134.

Barnes, A. and J. Parkhurst. 2014. Can global health policy be depoliticised? A critique of global calls for evidence-based policy. In *Handbook of Global Health Policy*, ed. G. Yamey and G. Brown, 157–173. Chichester: Wiley-Blackwell.

Biddle, J. 2013. State of the field: transient underdetermination and values in science. *Philosophy of Science* 44: 124–133.
Biller-Andorno, N., R. K. Lie and R. T. Meulen. 2002. Evidence-based medicine as an instrument for rational health policy. *Health Care Analysis* 10: 261–275.
Campbell Collaboration. 2016. *Methodological Expectations of Campbell Collaboration Intervention Reviews: Conduct Standards*. Campbell Policies and Guidelines Series No. 3.
Cartwright, N. and A. Marcellesi. 2015. EBP: where rigor matters. In *Foundations and Methods from Mathematics to Neuroscience: Essays Inspired by Patrick Suppes*, ed. C. E. Crangle, A. García de la Sienra and H. E. Longino. Stanford, CA: CSLI Publications.
Deaton, A. 2010. Instruments, randomization, and learning about development. *Journal of Economic Literature* 48: 424–455.
Douglas, H. 2009. *Science, Policy, and the Value-Free Ideal*. Pittsburgh, PA: University of Pittsburgh Press.
Duflo, E., R. Glennerster and M. Kremer. 2008. Using randomization in development economics research: a toolkit. In *Handbook of Development Economics*, vol. 4, ed. P. T. Schultz and J. Strauss. Amsterdam: North Holland.
Elliott, K. 2017. *A Tapestry of Values: An Introduction to Values in Science*. Oxford: Oxford University Press.
Elliott, K. and D. McKaughan. 2014. Nonepistemic values and the multiple goals of science. *Philosophy of Science* 81: 1–21.
Fink, G., M. McConnell and S. Vollmer. 2014. Testing for heterogeneous treatment effects in experimental data: false discovery risks and correction procedures. *Journal of Development Effectiveness* 6: 44–57.
Guyatt, G. H., A. D. Oxman, R. Kunz, J. Woodcock, J. Brozek, M. Helfand, P. Alonso-Coello, P. Glasziou, R. Jaeschke, E.A. Akl, S. Norris, G. Vist, P. Dahm, V.K. Shukla, J. Higgins, Y. Falck-Ytter and H.J. Scheunemann. 2011. Grade guidelines: 7. rating the quality of evidence – inconsistency. *Journal of Clinical Epidemiology* 64: 1294–1302.
Hausman, D. and M. S. McPherson, ed. 1996. How could ethics matter to economics? In *Economic Analysis and Moral Philosophy*, Appendix. Cambridge: Cambridge University Press.
Higgins, J. P. T. and S. Green. 2011. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0*, The Cochrane Collaboration, Retrieved 18 November 2017 from: http://handbook.cochrane.org.
Holland, P. W. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81: 945–970.
Hotz, V. J., G. W. Imbens and J. H. Mortimer. 2005. Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics* 125: 241–270.
Imai, K., D. Tingley and T. Yamamoto. 2013. Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society A* 176: 5–51.
Imbens, G. W. and J. M. Wooldridge. 2009. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 47: 5–86.
JPAL, Abdul Latif Jameel Poverty Action Lab. 2017. <https://www.povertyactionlab.org/methodology/how/how-obtain-results>.
Katikireddi, V. S. and S. Valles. 2015. Coupled ethical-epistemic analysis of public health research and practice: categorizing variables to improve population health and equity. *American Journal of Public Health* 105: e36–e42.
Kuhn, T. S. 1977. *The Essential Tension*. Chicago, IL: University of Chicago Press.
Longino, H. 1996. Cognitive and non-cognitive values in science: rethinking the dichotomy. In *Feminism, Science, and the Philosophy of Science*, ed. L. H. Nelson and J. Nelson, 39–58. Dordrecht: Kluwer.
Manski, C. F. 2000. Identification problems and decisions under ambiguity: empirical analysis of treatment response and normative analysis of treatment choice. *Journal of Econometrics* 95: 415–442.

Moher, D., S. Hopewell, K. F. Schulz, V. Montori, P. C. Gøtzsche, P. J. Devereaux, D. Elbourne, M. Egger and D. G. Altman. 2010. Explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *British Medical Journal* 340: c869.

Oxman, A. D. and G. H. Guyatt. 1992. A consumer's guide to subgroup analyses. *Annals of Internal Medicine* 116: 78–84.

Parkhurst, J. and S. Abeysinghe. 2016. What constitutes 'good' evidence for public health and social policy-making? From hierarchies to appropriateness. *Social Epistemology* 30: 665–679.

Pearl, J. 2014. Reply to commentary by Imai, Keele, Tingley and Yamamoto concerning causal mediation analysis. *Psychological Methods* 19: 488–492.

Rubin, D. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Education Psychology* 66: 688–701.

Tuana, N. 2013. Embedding philosophers in the practices of science: bringing humanities to the sciences. *Synthese* 190: 1955–1973.

Tuana, N., R. L. Sriver, T. Svoboda, R. Olson, P. J. Irvine, J. Haqq-Misra and K. Keller. 2012. Towards integrated ethical and scientific analysis of geoengineering: a research agenda. *Ethics, Policy and Environment* 15(2): 1–22.

Varadhan, R and J. D. Seeger. 2013. Estimation and reporting of heterogeneity of treatment effects. In *Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide*, ed. P. Velentgas, N. A. Dreyer, P. Nourjah, S. R. Smith and M. M. Torchia, 35–44. Rockville, MD: Agency for Healthcare Research and Quality.

Vedung, E. 2010. Four waves of evaluation diffusion. *Evaluation* 16: 263–277.

What Works Clearinghouse, Institute of Education Sciences, U.S. Department of Education. 2014. *What Works Clearinghouse: Procedures and Standards Handbook v.3.0*. <http://whatworks.ed.gov>.

What Works Clearinghouse. 2017. *What Works Clearinghouse: Procedures Handbook v.4.0*. Institute of Education Sciences, U.S. Department of Education. <http://whatworks.ed.gov>.

BIOGRAPHICAL INFORMATION

Donal Khosrowi is currently a final-year doctoral candidate in Philosophy at Durham University. His doctoral research focuses on the problem of extrapolating causal effects. His broader research interests are in causal inference, scientific representation and values in science.