Trade-offs between Epistemic and Moral Values in Evidence-Based Policy

Donal Khosrowi
Durham University
30 October 2016

Abstract: I examine the role and relationship of epistemic and moral values in the Evidence-Based Policy (EBP) paradigm. I argue that several epistemic values that play a crucial role in shaping standard EBP methodology stand in a trade-off relation with certain kinds of moral and political values. This is because the outputs afforded by standard EBP methods are insufficient for the pursuit of moral and political values that require information about the distribution of individual treatment-effects among agents in a population. I examine a potential reply to this standard concern, and argue that the changes to standard EBP methodology required for rendering research outputs informative about the distributive consequences of policy typically involve the sacrifice of several key EBP epistemic values at once. I expand on the implications of this trade-off for value-freedom and -neutrality in EBP.

Keywords: **Evidence-Based Policy, epistemic values, non-epistemic values, trade-off**

1. Introduction

Evidence-based policy (EBP) is the call that public policy formation should be informed by high-quality empirical evidence for policy effectiveness from randomized controlled trials (RCTs) and meta-analyses. In emphasizing the superior epistemic credentials of these methods, EBP advocates seek promote several epistemic values such as rigor, 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 that policy-makers may be interested in pursuing. Specifically, I argue that standard EBP methodology severely complicates policy makers' ability to pursue moral values such as equality or priority for the worst-off. This is because standard EBP methods are not informative about the *distributive consequences* of policy (see e.g. Manski 2000). This is a substantive shortcoming, particularly when we have reasons to suspect that a policy will render some agents worse off. Yet, since the evidence typically afforded by EBP methods is uninformative on such distributive consequences, it is differentially useful for the pursuit of different moral and political values, specifically utilitarian vs. non-utilitarian values. I argue that this challenges both value-freedom and neutrality in EBP.

The contents are organized as follows. In Section 2 I offer a sketch of the 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 standard EBP methodology faces with respect to generating information about the distributive consequences of policy from RCTs. I discuss how this problem can be addressed by performing subgroup analyses and expand on some of the challenges that this method faces. I also comment briefly on the extent to which these issues have been anticipated and addressed in the extant EBP literature. In Section 4 I give my argument for the trade-off relation between basic EBP epistemic values and moral values that are sensitive to the distributive consequences of policy. I expand on how this trade-off challenges both value-freedom and neutrality in the EBP paradigm. Section 5 concludes.

2. Epistemic Values in EBP

Before I sketch the central epistemic values in the EBP paradigm and how they relate to EBP methodology, it is important to note that there is perhaps no univocally 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]

Even so, it is not entirely misleading to think that there is a kernel of epistemic values that are common to a broad variety of activities under the EBP heading. It is this kernel of values that I focus on. 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 it seems more plausible to consider values that concern the production of treatment-effect estimates. More specifically, the values that I focus on are *rigor*, *unbiasedness*, *precision* and *the ability to obtain causal conclusions* on grounds of EBP evidence. I consider these values to be prima facie uncontroversial instances of purely epistemic values that seem to be shared among many EBP practitioners. While it is

[1] In addition to this caveat, it is important to note that the construal of Evidence-Based Policy I consider here is somewhat constrained in that it prevalently focuses on the so-called *treatment-effects literature* as instantiated in e.g. econometrics and evidence-based economics, evidence-based medicine and educational research. The distinctive characteristic of this literature is its predominant focus on experimental and quasi-experimental methods to estimate treatment effectiveness. 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 raising this important point about the scope of Evidence-Based Policy.

not always clear what these values specifically consist in, my argument is sufficiently broad to cover most plausible construals that they permit.[2]

The values that I focus on are central to EBP in the sense that they jointly give rise to (and are promoted by) standard EBP methodology, i.e. a set of salient methodological principles that seem to be shared among proponents of the paradigm.

For instance, EBP methodology specifically focuses on certain epistemic targets, i.e. causal conclusions about policy effectiveness. Moreover, EBP methodology is premised on principles concerning the relative desirability of certain kinds of evidence, e.g. by emphasis of the superiority of experimental and quasi-experimental contra 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 principles mediate between epistemic values and methods in the sense that EBP methodology promotes values such as rigor, unbiasedness and causal inference *in virtue of* recommending the use of RCTs.

2.1 Value Neutrality and Freedom in EBP

Aside from the identification of crucial EBP epistemic values, it is important to consider whether EBP involves some ideal of value-freedom and/or neutrality. Similar to the issue of identifying key EBP epistemic values, it is not obvious that EBP proponents in general pursue any specific ideal with respect to value-freedom and neutrality.

Even so, it seems that the EBP paradigm rests on a relatively broad axiological presupposition that a *division of labor* with regard to settling normative issues of what values policy should promote and settling factual issues of what are effective means to promote these values is possible. In other words, EBP proponents seem to assume that agreement on the desirability of policy outcomes can be *separated* from the production of evidence speaking for the efficacy and effectiveness of policy in realizing these outcomes.

---

[2] There may be several additional candidate epistemic values that appear to play prominent roles in shaping EBP research but are not considered here. One such candidate is *generality*, where the principled aim is to establish general claims about the causal efficacy of intervention-types that are robust across time, environments, populations, and individuals. This value seems particularly relevant for extrapolation of causal claims to novel targets; an issue that is related to, but epistemologically distinct, from the issue of welfare analysis of extant interventions that I focus on. I thank Heather Douglas for proposing this additional candidate value at the "Science, Values and Democracy" workshop in Tilburg, NL.

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 policy makers have settled issues concerning the relative desirability of social outcomes (cf. Hausman and McPherson 1996). While I am not claiming that EBP proponents subscribe to this particular ideal, EBP methodology seems to presuppose at least that *some such* division of labor is possible. Let me expand on what this suggests for the role of value-freedom and neutrality in EBP.

First, it seems plausible that many EBP proponents pursue 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*. For instance, while non-epistemic values may be involved in selecting outcome variables of interest, or may act as constraints on whether conducting RCTs is morally permissible, non-epistemic values are generally not and should not be involved in the choice and application of methods once these issues are settled. For instance, the choice between RCTs and observational studies, or the interpretation of estimands obtained from such studies, should not vary with respect to researchers' preferred conclusions about the desirability of the policies under scrutiny. These *internal* aspects should be guided by epistemic values alone.

Second, I consider EBP proponents to pursue *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 about the relative desirability of social outcomes or the interventions that promote them. 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-neutral, the adequacy of such presuppositions speaking for the desirability of some social outcome must be settled *independently* from (and perhaps prior to) generating information about the relative effectiveness of different interventions in producing the outcome. If such independence is achieved, then *even if* EBP research sometimes issues normative claims, these claims are still value-neutral since they remain non-committal on the adequacy of the substantive moral value presuppositions involved. This issue is left to policy-makers to settle.

With this brief exposition in mind, let me focus on the underlying reasons for why the epistemic challenges involved in generating information about the distributive consequences of policy yield a trade-off between the epistemic values outlined above and non-epistemic values such as equality and priority for the worst-off.

3. Treatment Effect Heterogeneity

Public policy interventions almost invariably affect agents in heterogeneous ways. Consider for instance the case of *microfinance programs*, i.e. programs that supply microcredits to agents who lack access to capital markets. Let us grant for the moment that at least some of these programs may be successful in generating positive long-run welfare consequences for target populations, e.g. by increasing average household endowment or private investment. Even so, behavioral response to microfinance access often differs significantly between agents (cf. Banerjee et al. 2015)[3]. Some agents, e.g. those whose otherwise successful entrepreneurial efforts are inhibited by inadequate access to capital markets, may significantly benefit from such programs. 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 program loans.

Such heterogeneity in individual treatment effects is predominantly 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. This means that the mechanisms connecting treatment and outcome variables of interest typically involve various factors other than treatment that affect the causal relations between treatment and outcome in different ways. For instance, the mechanisms that causally relate microfinance access and eventual welfare consequences for target agents are plausibly mediated and moderated by an extensive battery of factors such as entrepreneurial ability, education, prior business ownership, pre-intervention budget constraints, business plan feasibility etc. These and other factors jointly moderate or mediate the causal effect of treatment on outcome, and agents will typically differ with respect to their individual-specific realizations of these factors as well as whether and how these factors are involved in the individual-specific mechanism that govern the production of the outcomes of interest. As a consequence of such differences, individual treatment effects with respect to one and the same intervention will typically differ between individuals.

This kind of causally relevant 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 policy outcomes may differ significantly between industries, individual firms and other agential units; and in public health and development economics, where agents' response to programs such as bednet distribution might exhibit substantial heterogeneity as a function of agents' basic needs or epidemiological knowledge.

[3] Cited with permission from the authors

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 marked contrast to evidence-based medicine, see e.g. Oxman and Guyatt 1992 for an early treatment). This is surprising because heterogeneity is 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

Technically, treatment effect heterogeneity is the systematic variation in the sign and/or magnitude of individual treatment effects among agents 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.

RCTs can be considered to remedy this inferential dead-end at least to some extent by permitting the estimation of *average treatment effects* (ATEs) instead of ITEs. This is achieved by randomization of confounding factors and treatment moderators and mediators[4] through 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) are successful in that the net effects of confounders and moderators (as well as their interactions) are approximately balanced between treatment and control groups, an ideal RCT can help obtain a consistent 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 ATE, however, does not permit inferences about ITEs. At best, and in the absence of any knowledge about treatment effect covariates such as moderators and mediators 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 (suspected) heterogeneity among treatment-effect covariate realizations and consequently ITEs, this estimate will not be

[4] The distinction between confounders and moderators/mediators being that confounders influence the outcome variable independently of treatment whereas moderators/mediators influence the outcome by affecting the causal pathway(s) connecting treatment and outcome.

precise, so accurate inferences about ITEs are largely precluded and information on heterogeneity cannot be recovered from $\widehat{ATE}$.[5]

This has significant bearing on the *transferability* of trial results, i.e. the extent to which the ATE from a study population A can be expected to be replicated in some other population B. Two jointly sufficient conditions for the transferability of trial results 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 claim to be extrapolated. The second condition is that the distribution of treatment effect covariates in the target is the same in both populations (see e.g. Cartwright and Marcellesi 2015 for similar conditions).[6] 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 effected by differences in treatment-effect covariates as well as how such covariates such as moderators and mediators 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-Imbens and Mortimer 2005, Duflo, Glennerster and Kremer 2008; Imbens and Wooldridge 2009; Bareinboim and Pearl 2013; Cartwright and Marcellesi 2015).

However, 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 *any* policy formation process that is concerned with the *distributive consequences* of policy (cf. Manski 2000). More specifically, policy-makers are often 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 heterogeneous treatment effects are distributed among agents, with respect to which observable baseline characteristics, whether heterogeneity obtains in magnitude or also in sign, etc.

This information is crucial particularly in those cases where it is reasonable to suspect that at least some agents may respond negatively to 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 policy that will render some agents worse off? If so, how

[5] While this difference-in-means estimation yields, without strong assumptions, an unbiased estimate of the sample ATE, and under somewhat stronger assumptions of the population ATE, it takes substantive assumptions about distributions of ITEs to estimate even the sample variance of the ATE (although this estimate can be bounded by inspection of the treatment and control mean variances).

[6] Necessary conditions might be weaker, cf. Bareinboim and Pearl (2013)

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

As these stylized concerns suggest, policy-makers may be interested in pursuing a variety of different distributive values. Yet, in order to pursue these values rigorously, in the sense that they have good reasons to believe that an intervention will promote them, policy-makers 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 do not provide such information on their own.

Yet, this does not mean that EBP methodology is at a complete loss in this regard, as EBP proponents may be keen to point out that one way to address this problem is to perform so-called *subgroup analyses*. However, I argue below that performing such analyses comes at the expense of sacrificing several key EBP epistemic values and that this creates a tradeoff between the 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, Glennerster and Kremer (2008), subgroup analyses partition experimental populations into subgroups according to observable 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 run on the partitioned data to obtain conditional, subgroup-specific ATEs (CATEs). An alternative to this stratification approach that is applied predominantly when investigating binary and categorical variables, is to run so-called *meta-regressions*, where potentially interesting treatment-effect covariates are modeled as interaction terms with treatment in a standard regression framework. In doing so, it is possible to obtain information on significant interaction effects between observables and treatment that may be taken as evidence for the involvement of the respective treatment effect covariates as moderators or mediators.

Even so, while subgroup analyses seem to offer 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 realization of EBP epistemic values.

First, the information that meta-regressions can generate is purely correlational in nature, and hence subject to standard concerns about endogeneity and consequent bias. For instance, statistically significant parameter estimates on treatment effect heterogeneity of microfinance programs with respect to differences in prior business ownership do not permit the straightforward interpretation that prior business ownership is a causally relevant treatment effect covariate.

This is because the significance of the estimate may be attributable to common-causes, e.g. because business ownership is highly correlated with business education, and it is business education that is causally relevant for the production of microfinance outcomes, but prior business ownership in the absence of business education may not contribute at all to outcomes of interest.[7] 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 biased.

More generally, parameter estimates for treatment effect covariates will invariably remain subject to such concerns about bias unless we can entertain the relatively strong assumption that the regressors are uncorrelated with the error term of the meta-regression (see e.g. Pearl 2014). However, it is precisely such assumptions, which are necessary for unbiased identification in regression contexts, that EBP proponents are typically keen to avoid and that are expressly dismissed in the methodological tenets that emphasize randomization as the key strategy to avoid questionable identification assumptions.

Randomization at the treatment stage does not alleviate these concerns either, because treatment effect moderators are not necessarily randomly distributed among agents who, with respect to one subgroup characteristic, may *systematically* differ on several other relevant and collinear or interacted covariates at once. This means that obtaining *unbiased estimates* and straightforward *causal conclusions* about the role of covariates as treatment moderators is typically precluded, threatening at least two EBP epistemic values at once.

A second worry about subgroup analyses concerns the *precision* of effect estimates and *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

[7] 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 covariate realizations (see e.g. Imai et al. 2013). While such designs are in principle feasible, they also raise issues with precision and statistical power.

occurrence of spurious results exceedingly likely. At the very least, suitable statistical corrections for multiple hypothesis testing 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, Glennerster and Kremer 2008, 65), this is rarely carried out in practice (cf. Fink et al. 2014, 47). Moreover, to alleviate concerns about insufficient statistical power and precision, sample sizes may need to be expanded for subgroup analyses to be sufficiently informative. For instance, in order to detect a heterogeneity signal of the same magnitude as the ATE and with the same precision as the 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 are often significantly smaller than ATEs, which requires much greater expansions of sample size to maintain sufficient power.

These and other, related concerns severely limit the extent to which subgroup analyses can inform about treatment effect heterogeneity. At most, and in line with standard recommendations (e.g. Varadhan and Seeger 2013), subgroup findings should 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 trials on potentially vulnerable subgroups appears to be a viable strategy to address 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 treatment effect covariates are, how they affect the outcomes of interest as well as which agents exhibit beneficial vs. harmful realizations of such covariates. So precise information on heterogeneity is still required even if we are willing to conduct subsequent trials on vulnerable subgroups.

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 *between-trial* heterogeneity, which is a related but conceptually distinct issue from the *within-trial* and *between-subject* heterogeneity that I consider here.

Let me expand on how these epistemic challenges for informing about heterogeneity create a trade-off between epistemic and moral values 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. broadly utilitarian and non-utilitarian values respectively.

Standard EBP methods such as RCTs, Regression Discontinuity Designs and IV identification strategies are in general capable of generating outputs that are sufficient for the pursuit of standard utilitarian values, i.e. those that are concerned with the increase or maximization of aggregate or average welfare. This is because the distribution of individual-specific contributions to aggregate welfare outcomes is not a primary concern for increasing aggregate or average welfare, so information on heterogeneity is not necessary for the pursuit of these values.[8]

Yet, such 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 initial distribution of welfare among agents as well as information on the changes to this distribution brought about by the intervention at issue. Yet, as I have argued above, such information on treatment effect heterogeneity 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. Moreover, methods such as Causal Bayes Net Analysis, Qualitative Comparative Analysis, Process Tracing and Machine Learning may present potentially superior alternatives for the identification of causally relevant treatment effect covariates that generate heterogeneity. However, such techniques are rarely acknowledged or mentioned in the standard manuals circulating in the EBP literature (e.g. Angrist and Pischke 2009), and even if they were, these methods are often neither straightforwardly compatible with the identification strategies that EBP practitioners typically pursue nor with the evidence ranking schemes that EBP methodologists subscribe to.

This licenses two conclusions. First, EBP methodology presently favors the production and use of evidence suitable for the pursuit of utilitarian values, i.e. those that focus on increasing or maximizing average or aggregate welfare. Second, EBP methodology presently fails to adequately promote or even hinders the production of high-quality evidence on heterogeneity that is necessary for the pursuit of many non-utilitarian values. As a consequence, standard EBP methodology renders the pursuit of distributive values such as egalitarian or prioritarian ones relatively more difficult or infeasible.

[8] It might still be helpful, since welfare *maximization* is easier to accomplish when we have information that helps pick out those individuals who will likely benefit most from some intervention; granted that interventions can be targeted to affect only such individuals.

This generates a trade-off between the epistemic values central to EBP and the moral and political values that policy-makers are in a position to pursue effectively on grounds of EBP evidence. More specifically, whenever the pursuit of moral and political values requires information on distributive consequences of policy, standard EBP evidence fails to provide the required information. Conversely, whenever evidence of the kind required to inform about distributive consequences of policy shall be produced, this requires at least some sacrifice of basic EBP epistemic values. More specifically, whenever EBP methodology and methods are changed in order to generate information on heterogeneity, e.g. by means of subgroup analyses, this comes at the expense of sacrificing at least three crucial EBP epistemic values at once, i.e. the *unbiasedness* and *precision* of effect estimates, as well as the *ability to obtain causal conclusions*. Maintaining these values, on the other hand, comes at the expense of sacrificing the informativeness of EBP research outputs about the distributive consequences of policy.[9]

Let me expand on what this trade-off implies for value-freedom and neutrality in EBP. First, if the value-free ideal underlying the EBP paradigm is to say that non-epistemic values are generally not and should not be involved in shaping the conduct and outcomes of EBP research internally, then the desirability of this ideal is challenged. The reason is that moral and political values are at least involved to the extent that without suitable changes to EBP methodology, the pursuit of non-utilitarian values is inhibited. If this situation should be remedied, then this requires changes to methodology that privilege or prioritize the production of evidence on heterogeneity. However, and this is the crucial point, these changes will be *effected by moral values*, since it is the pursuit of moral values that motivates the requisite changes to methodology. To the extent that these changes to methodology are justifiable and justified, this means that value-freedom in EBP is not a desirable ideal, even at internal stages such as method choice and model specification.

Value-neutrality is challenged as well. It assumes that once the desirability of some social outcome is agreed upon, evidence speaking in favor of the effectiveness of some intervention in realizing this outcome at most figures in conditionally normative policy recommendations.

---

[9] 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). However, my point is weaker than Longino's in the sense that it 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 as they exhibit a demonstrably political (or moral) valence (ibid.). Instead, even if we grant that these values 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.

Yet, inferences about policy effectiveness are typically grounded in information about ATEs and as such do not accommodate information on distributive consequences. So this way of operationalizing what it means for a program to be effective brackets concerns about heterogeneity. As it stands, an *effective* program is considered a good program 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 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.

To maintain neutrality with respect to distributive values it is not enough to agree on the desirability of social outcomes *as such*. It is also necessary to agree upon the *ways in which* these outcomes may be realized, since 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, some of which may be more or less desirable *in themselves*. This issue is masked when broadly utilitarian values are pursued, but becomes apparent when distributive consequences matter; as is the case for the pursuit of egalitarian and prioritarian values. So if we 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 policy and might be misleading about what *effective* programs are ultimately able to do for us, given the specific moral and political values that we pursue.

So at present, it seems that the dissemination of EBP research is premised on the implicit value presupposition that the relevant magnitude for deciding which policy to implement is its effectiveness in terms of average treatment effects. And this fails to be value neutral in the envisioned sense because it assumes that average effectiveness is the proper target of interest rather than delegating the question of whether it is, to policy makers and other agents to settle. In a nutshell, in order to maintain a traditional ideal of value-neutrality, additional value presuppositions such as the above must be made explicit for EBP policy recommendations to remain value-neutral in the envisioned sense.

5. Conclusion

I have argued that there exists a trade-off relation between key EBP epistemic values and non-epistemic values that are sensitive to distributive consequences of policy, 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 ideals of value-freedom and neutrality in the EBP paradigm. This may be taken as starting point to reconsider some of the standard epistemic value presuppositions entertained in EBP as well as for refining

EBP methodology in ways that enable and facilitate the pursuit of a wider range of moral and political values.

References

Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. "Mostly harmless econometrics." Princeton: Princeton University Press.

Banerjee, Abhijit, Emily Breza, Esther Duflo, and Cynthia Kinnan. 2015. "Do Credit Constraints Limit Entrepreneurship? Heterogeneity in the Returns to Microfinance". Working paper.

Bareinboim, Elias, and Judea Pearl. 2013. "A General Algorithm for Deciding Transportability of Experimental Results." Journal of Causal Inference. 1: 107-134.

Cartwright, Nancy, Alexandre Marcellesi. 2015. "EBP : Where Rigor Matters." In Foundations and Methods from Mathematics to Neuroscience : Essays Inspired by Patrick Suppes, ed. Colleen E. Crangle, Adolfo García de la Sienra, Helen E. Longino, Stanford: CSLI Publications.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2008. "Using Randomization in Development Economics Research: A Toolkit." In Handbook of Development Economics, Vol. 4, ed. Paul T. Schultz and John Strauss. Amsterdam and New York: North Holland.

Fink, Günther, Margaret McConnell, and Sebastian Vollmer. 2014. "Testing for Heterogeneous Treatment Effects in Experimental Data: False Discovery Risks and Correction Procedures." Journal of Development Effectiveness. 6:44-57.

Hausman, Daniel, and Michael S. McPherson. 1996. "How Could Ethics Matter to Economics?" In Economic Analysis and Moral Philosophy, Hausman and McPherson, Appendix. Cambridge: Cambridge University Press.

Holland, Paul W. 1986. "Statistics and Causal Inference." Journal of the American Statistical Association. 81:945-970.

Hotz, V. Joseph, Guido W. Imbens, and Julie H. Mortimer. 2005. "Predicting the efficacy of future training programs using past experiences at other locations." Journal of Econometrics. 125:241–270.

Imai, Kosuke, Dusting Tingley, and Teppei Yamamoto. 2013. "Experimental designs for identifying causal mechanisms." Journal of the Royal Statistical Society A, 176 Part 1:5-51.

Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." Journal of Economic Literature. 47:5-86.

Kuhn, Thomas S. 1977. *The Essential Tension*. Chicago, IL: University of Chicago Press.

Longino, Helen. 1996. *Cognitive and Non-Cognitive Values in Science: Rethinking the Dichotomy, in Feminism, Science, and the Philosophy of Science*, ed. Lynn Hankinson Nelson and Jack Nelson, 39-58. Dodrecht: Kluwer.

Manski, Charles F. 2000. "Identification problems and decisions under ambiguity: empirical analysis of treatment response and normative analysis of treatment choice." Journal of Econometrics. 95(2):415–442.

Oxman, Andy D., and Gordon H. Guyatt. 1992. "A Consumer's Guide to Subgroup Analyses." Annals of Internal Medicine. 116:78–84.

Pearl, Judea. 2014. "Reply to Commentary by Imai, Keele, Tingley and Yamamoto Concerning Causal Mediation Analysis." Psychological Methods. 19(4):488-492.

Rubin, Donald. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." Journal of Education Psychology. 66:688-701.

Varadhan, Ravi, and John 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. Patricia Velentgas, Nancy A. Dreyer , Parivash Nourjah, Scott R. Smith, Marion M. Torchia, 35-44. Rockville, MD: Agency for Healthcare Research and Quality.