Evidence and Evidence-Based Policy

Antony Eagle

Choices, Models and Morals » Lecture 7

Contents

Making Policy: Designing Interventions

Treatments and Policies

- A **treatment** is an intervention (perhaps including just 'leaving things be') intended to produce a net beneficial change.
  - In the **policy** context, a treatment might be some policy designed to improve wellbeing, at a given monetary cost. In **medicine**, the benefit is to health, and the costs are side-effects.
- A decision to treat, or implement policy, can be modelled using the decision-theoretic framework we've already discussed in lecture 2. Namely, we are advised to choose the treatment that maximises expected net benefit.

> the notion of evidence-based policy and practice '… fits well with a rational decision-making model of the policy process'.… Thus, it appears to be rational common sense to see policy as a purposive course of action in pursuit of objectives based upon careful assessment of alternative ways of achieving such objectives and effective implementation of the selected course of action (Sanderson 2002: 5)

- Even accepting this framework, it is crucial that we can identify **what effects our acts could produce** – whether with certainty, or with some degree of probability.
- In the policy context, **economic theory** is supposed to give us information about the causal mechanisms governing the rational interactions between economic agents that drive policy outcomes.

Public Policy and Economic Theory

- In lecture 5 we discussed the question of whether economic models *explain*.
- This was a pressing question, raised by an inconsistent triad of plausible claims (Reiss 2013: 127): (i) only true theories explain; (ii) most economic models are false of/inaccurate about their target system; (iii) economic models are explanatory. These can be generalized:
  (1) Only true causal claims can be the basis of effective policy design;
  (2) Most economic models, and the theories using them, are false of/inaccurate about their target systems;
  (3) Economic theories are central to the design and evaluation of policy.
- Reiss (2013: 127–33) gives an extensive defence of (2).
- In practice, (3) seems true; the policies governments design are sold to us on the basis of their cost-effectiveness at promoting some desired end, based generally on some underlying rational choice model of the effects of the policy intervention.

Economics Without Economic Theory

- A recent development is the rise of **evidence-based policy** which aims to provide a way out of this trilemma by denying (3).
- The concept was developed in medicine, where there is a distinct lack of good theories.
  - We have pretty good theories of basic biochemistry, but at the level of the human organism, nothing that looks like theory (Reiss 2013: 198).
- The idea is to try to reach conclusions about **what works** without needing to derive them from a theory of *how* and *why* it would work.
  - So evidence-based approaches are supposed to be **contrasted** with 'mechanism-based reasoning' (Howick, Chalmers, *et al.* 2011). The latter of course can be evidence-based, as it is in natural sciences, where inference to mechanisms and causes of the basis of experimental evidence is standard.
- The challenge posed by false economic models might be **side-stepped** by adopting an evidence-based approach in economics too.
  - Causal inference in econometrics requires substantive model-specific assumptions about causal mechanisms (the distinction between exogenous and endogenous variables) – if we don't have good justification for these assumptions, how can we base policy on them (Reiss 2013: 201–2)?

The Situation in Medicine

- In the medical context, we have the perfect conditions for **epistemic disaster**:
  - A highly emotive and high stakes subject matter, with vulnerable people seeking treatment;
  - A highly complex system without good knowledge of the mechanisms driving it, and a large role for noise/chance;
  - Potentially significant money to be made from selling treatments;
  - A long established history of 'tried and true' treatments, both medical and 'alternative'/'complementary' with significant entrenchment effects;
  - A cult of 'expertise' around doctors and alternative health practitioners.
- It is unsurprising that **disinformation** flourishes (Evans, Thornton, *et al.* 2011). Patients desperate for a cure confront a health landscape populated with snake oil salesmen, paternalistic doctors keen to protect their turf, old wives tales, and cognitive biases that attribute causation to mere happenstance ('Nature cures, but the doctor takes the fee' Evans, Thornton, *et al.* (2011), p. 64).

Evidence-Based Medicine

> the triumphs of modern medicine can easily lead us to overlook many of its ongoing problems. Even today, too much medical decision-making is based on poor evidence. There are still too many medical treatments that harm patients, some that are of little or no proven benefit, and others that are worthwhile but are not used enough. How can this be, when every year, studies into the effects of treatments generate a mountain of results? Sadly, the evidence is often unreliable and, moreover, much of the research that is done does not address the questions that patients need answered.
> Part of the problem is that treatment effects are very seldom overwhelmingly obvious or dramatic. Instead, there will usually be uncertainties about how well new treatments work, or indeed whether they do more good than harm. So carefully designed fair tests – tests that set out to reduce biases and take into account the play of chance… – are necessary to identify treatment effects reliably.
> (Evans, Thornton, *et al.* 2011: xx)

Evidence-Based Medicine in Practice

- Central to the practical implementation of a rational choice approach to treatments is the idea of an **evidence hierarchy**, where different types of evidence are taken to be more or less reliable in justifying causal beliefs – see table 1.
- This is supposed to influence clinical practice: 'Good decisions should be informed by good evidence, which will tell us about the likely consequences of different treatment options' (Evans, Thornton, *et al.* 2011: 143).

Table 1: Levels of Evidence Hierachy: Treatment Benefits (Howick, Chalmers, *et al.* 2011).

Quality of evidence	Type of evidence
Step 1	Systematic review of randomized trials or n-of-1 trials
Step 2	Randomized trial or observational study with dramatic effect
Step 3	Non-randomized controlled cohort/follow-up study
Step 4	Case-series, case-control studies, or historically controlled studies
Step 5	Mechanism-based reasoning

Mill's Method of Difference

- Why are RCTs at the top of the hierarchy?
- A basic account of causal inference in experimental conditions is provided by Mill's **method of difference**:

> if an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring in the former; the circumstance in which alone the two instances differ, is the effect, or the cause, or an indispensable part of the cause, of the phenomenon.…
> The Method of Difference has for its foundation, that whatever can not be eliminated, is connected with the phenomenon by a law. Of these methods, that of Difference is more particularly a method of artificial experiment …
> It thus appears to be by the Method of Difference alone that we can ever, in the way of direct experience, arrive with certainty at causes. (Mill 1874: bks III, ch. VIII, §§2–3)

Applying Mill's Method

- Mill's method says that if we can **control** for other independent sources of an effect, and through varying just one factor thereby produce a difference in the effect, we have identified that factor as a cause.
  - We have given ourselves good grounds to endorse this counterfactual: '*If C hadn't occurred, E wouldn't have occurred*', and that counterfactual is sufficient for causation on many views (Lewis 1973).
- However it is very difficult to apply this method: particular instances differ in ever so many ways, so that we generally cannot control for every potentially confounding factor in a specific case.
  - If establishing causal efficacy of a medical treatment requires holding everything else fixed, then we cannot establish whether a drug cures a disease, since mostly we cannot find genetically identical individuals to use as controls.

Randomised Controlled Trials

- A way around this is to try and establish an **aggregate** effect: not comparing the difference between two **instances**, but between two **populations**, one of which is subject to treatment and the other of which is not.
- If the populations are well matched in other respects, then the only **net difference** between them is the treatment. If there is a significant difference in outcome, we can identify the treatment as a cause (Cartwright 2008: 129–30).
- Technically, we need that the other **confounding factors** (potential causes) are **identically distributed** between the control and treatment groups (Cartwright 2007: 16; Reiss 2013: 201).
- This is normally secured by **randomizing** allocation to the treatment and control groups; then the allocation has not been determined by any factor that will then turn out to be common to all members of the treatment group.
  - If you randomise, **by chance** you may not achieve an identical distribution.
- We need also to identify what a **significant** difference is; because there are doubtless other unknown confounding factors present in both subpopulations, so we shouldn't expect to see all and only members of the treatment group exhibit the effect.

RCTs and Policy Evaluation

- For policy-makers who want to know whether a certain policy intervention should be adopted, given an antecedently set policy objective, an RCT provides an important **part** of the story, since it will tell you whether the intervention causes the desired outcome.
- But an RCT will not tell you:
  1. Whether you **ought** to have that policy objective (Sanderson 2002: 4) (how to assign values to outcomes);
  2. Whether a **different intervention** would be more effective (Sanderson 2002: 4) (that is the role of comparing treatments for expected efficacy);
  3. **How** an intervention yields its effect – it provides no knowledge of **mechanisms** (Sanderson 2002: 18).
- Often, of course, circumstances are such that evaluation of whether an intervention would work via an RCT is almost the entirety of what policy-makers can evaluate.
  - Often policy makers have no control over objectives (which are set exogenously, perhaps by political masters); and
  - They often only have a limited palette of possible interventions (other effective interventions being ruled out as too costly, unethical, or politically infeasible).

Limitations of RCTs

What works may not be seen to work

- In practice many policy evaluations don't involve an RCT at all.
- For example, many policies are justified by **pilot studies**, but these may not reflect the policy-as-implemented:

> The first problem concerns the time needed for the effects of new policies to be manifested … It may take some considerable time for pilot projects to become fully established so as to represent the conditions under which a policy would work when fully implemented. If the policy aims to change attitudes and behaviour or achieve institutional reform, effects may be difficult to achieve during the course of a pilot project.
> This problem is exacerbated by political considerations that constrain the length of time for pilot programmes to operate. When policy initiatives arise from political manifesto commitments, policy makers are understandably impatient to receive results that will provide evidential support for decisions to proceed with full implementation (Sanderson 2002: 11)

Hurdles to Methodology

- Ideal RCTs can yield causal information. But can we even implement a properly designed RCT?
  1. The method assumes identically distributed confounders. But in small populations, like those pilot studies, it can be almost guaranteed that potentially confounding effects are not identically distributed (Reiss 2013: 203).
  2. Awareness of treatment is a significant confounder. In drug trials, it is often possible to make sugar pills that are indistinguishable from the drug. But it is very hard to disguise from participants in policy trials that they are subject to a particular policy (Reiss 2013: 204).
  3. Recruitment of subjects might be biased because RCTs have rigorous protocols; Reiss (2013: 204) worries about the 'risk-averse' not participating; consider also a trial of a 'cashless welfare card' that moves some people off cash payments – the most vulnerable may be disinclined to do anything to jeopardise an already precarious existence, hence any positive treatment effect may derive from the fact that only the 'quite poor', rather than the 'very poor', took up the trial.

What works may not be permitted to be tested

- A well-designed RCT will provide evidence of causal efficacy of a given intervention, but there are also concerns about whether it is possible or permissible to implement an RCT to test an intervention.
- It may not be **possible** to construct an RCT because the available experimental subjects may not be sufficiently **randomizable**.
  - For example, suppose we are conducting a trial of some public health intervention. We need a cohort of sick individuals, and since we cannot deliberately infect anyone, we have to work with what we have. But if there is some known confounder that some of the sick have, we face a choice: exclude those subjects so the population doesn't match the experimental cohort; or include them without randomising their allocation to treatment.
- It may not be ethically **permissible** to construct an RCT (Sanderson 2002: 12).
  - In a welfare trial, we cannot withhold welfare from some people just to see the effects of forcing people to 'fend for themselves'. In a democratic state, everyone has a vote: we cannot test the outcomes of other electoral methods if that involves denying people the right to vote.

What works here may not work there

- A key limitation of RCTs is that they detect causation in one **fixed population**, with certain frequencies of confounders, and certain underlying traits.
- What we need for policy design is some evidence that this intervention will **continue** to be effective in new circumstances.

> The methods recommended by typical evidence-ranking schemes [i.e., RCTs] are very good at establishing *efficacy*: whether a treatment causes a given outcome in the selected population under the selected circumstances. In evidence-based policy we are interested in *effectiveness*: What would happen were the treatment to be introduced as and when it would be in the population of interest. How can we move from efficacy to effectiveness? (Cartwright 2008: 130–31)

The Case of California Class Sizes

> there are a number of other reasons why a pilot may not be typical of the policy as it would ultimately be implemented. … as Hasluck … points out, '… the resources devoted to the pilot may exceed that (sic) available at national roll out. There may also be a greater level of commitment and a "pioneering spirit" amongst staff involved in delivery' (Sanderson 2002: 12)
> Consider the California class-size reduction programme. The plan was backed up by evidence that class-size reduction is effective for improving reading scores from a well-conducted RCT in Tennessee. Yet in California when class sizes were reduced across the state reading scores did not go up. … There's a conventional explanation. … California rolled out the programme state-wide and over a short period creating a sudden need for new teachers and new classrooms. So large numbers of poorly qualified teachers were hired and not surprisingly the more poorly qualified teachers went to the more disadvantaged schools. Also classes were held in spaces not appropriate and other educational programmes commonly taken to be conducive to learning to read were curtailed for lack of space (Cartwright 2008: 131; see Reiss 2013: 205–6)

The Curse of Causal Agnosticism

> RCTs, when implemented successfully, give us knowledge "cheaply" in the sense that they require no specific background knowledge in order to identify a causal effect from the data. But this does come at an eventual cost: if the understanding of the causal structure that is being experimented on in the RCT is very limited, there are no good grounds for believing that a result will also hold in this or that population that differs from the test population.
> In a sense an RCT is a "black-box" method of causal inference. A treatment is administered, an outcome observed, with no need for any understanding of what is going on in between and why a treatment produces its outcome. But if there is no knowledge of why a treatment produces a given outcome, the basis for drawing inferences beyond the immediate test population is very narrow. (Reiss 2013: 205)

The Problem of Reflexivity

- In many policy cases, the mechanisms of the policy are intended to influence people's actions and choices.
- An RCT is good in detecting **fixed causal structure**. But if people change their beliefs or desires in response to policy changes, then the results of the RCT won't apply, since the **causal environment** will have changed.
  - This may even undermine the RCT itself, as has been argued for certain medical RCTs, where the subjects access to their own health improvements will bolster any placebo effects, meaning that causally relevant factors will not be fixed across treatment and control groups (Chemla and Hennessy 2019).
- The problem is that people can **foresee** the policy implications, and they act by responding **strategically to the policy**.
  - An RCT gives you information about how your policy will affect things given an unknown state of nature. But if the unknown state includes people who are responding rationally to your policy choices, then your actions may not have their desired consequences: it will be like playing a prisoner's dilemma where you play based on the probabilities, and the other player plays strategically.

Policy Evaluation and Inductive Risk

Causal Conclusions and Significance

- The difference between control and treatment group in an RCT will typically be **noisy**: you will see some effects in both groups, due to random chance in the composition of those groups.
- The causal conclusion will depend on whether an observed difference is **statistically significant**.
  - An observed difference in an RCT is significant at the p-level if the chance of observing that difference between treatment and control groups, given the **null hypothesis** that there is **no causal efficacy** to the intervention, is less than p.
  - Typical p-values used in significance testing are 0.05, 0.01.
- A **false positive** (or **type I error**) occurs when a significant result is observed even though the null hypothesis is correct.
- A **false negative** (or **type II error**) occurs when an insignificant result is observed even though there is causal efficacy to the intervention.
- Choices of p-value affect the error rates: set p low, and you raise the frequency of false negatives; set p high, and you raise the frequency of false positives.

Policy Implications and p-values

- If the question is purely scientific, the risks of type I/type II errors are **epistemic**: the risks of false belief, what Hempel (1965) calls 'inductive risk'.
- But if the results of the question inform practical decision making – for example, if a significant result in an RCT will determine an intervention that will impact people — then the risks of type I/type II error may have **welfare costs**.
  - A false positive may mean costly expense implementing an ineffective policy.
  - A false negative may mean opportunity costs as an effective policy is foregone.

The Case of Dioxins

> The deliberate choice of a level of statistical significance requires that one consider which kind of errors one is willing to tolerate. … In testing whether dioxins have a particular effect or not, an excess of false positives in such studies will mean that dioxins will appear to cause more harm to the animals than they actually do, leading to overregulation of the chemicals. An excess of false negatives will have the opposite result, causing dioxins to appear less harmful than they actually are, leading to underregulation of the chemicals. Thus, in general, false positives are likely to lead to stronger regulation than is warranted (or overregulation); false negatives are likely to lead to weaker regulation than is warranted (or underregulation). Overregulation presents excess costs to the industries that would bear the costs of regulations. Underregulation presents costs to public health and to other areas affected by damage to public health. Depending on how one values these effects, an evaluation that requires the consultation of non-epistemic values, different balances between false positives and false negatives will be preferable (Douglas 2000: 566–67)

But RCTs are better than nothing

- Type I/II errors are less likely to show up in **multiple** RCTs.
- Hence the recent emphasis in policy circles on **meta-analyses**, where many RCTs are aggregated to see the overall **pattern** of causes.
- These can be powerful tools in policy evaluation. Consider the discovery that when children who commit offences are formally processed by the legal system (rather than diverted to social programs or just released), this actually leads to increased subsequent criminality:

> juvenile system processing appears to not have a crime control effect, and across all measures appears to increase delinquency. … Given the additional financial costs associated with system processing (especially when compared to doing nothing) and the lack of evidence for any public safety benefit, jurisdictions should review their policies regarding the handling of juveniles (Petrosino, Turpin-Petrosino, and Guckenburg 2010: 6–7)

- For more examples of effective/ineffective policies, see 80000hours.org/articles/can-you-guess/.

Evidence: Measuring Economic Phenomena

Gathering Evidence

- To deploy an RCT, we need to determine what the effect of treatment is.
- In the simplest cases, this is straightforward: the patient recovers or they do not.
- But in most health policy scenarios, the treatment variable is not a simple **observable**.
- We might be interested in **quality of life**, not mere existence, after a health intervention.
- In economic policy, we might be concerned with the impact of a new policy on some economic variable of interest.

> Consumer price inflation will be my main case but I will briefly discuss GDP and unemployment as comparisons. All three variables are regarded as observable by economists. But, as we will see, measuring them requires making a large number of substantial, and often contentious, assumptions. Making these assumptions requires real commitment on the part of the investigator as regards certain *facts* relevant to the measurement procedure, the *measurement purpose* as well as *evaluative judgments*. (Reiss 2013: 150)

The Theory-Dependence of Observation

- I am principally concerned with the need for a commitment to prior facts, which are often supplied by a theoretical framework the investigator adopts.
- This is typically labelled the **theory-ladenness of observation** (Boyd and Bogen 2021). As Hanson puts it, 'there is more to seeing than meets the eyeball' (Hanson 1958: 7).
- The boldest version of this idea is that our theories might influence our **perceptions**.
- This is the idea that perception is **penetrable**: that the concepts we have, or the theories we accept, shape the contents of perception.
  - In commenting on the history of science, Kuhn makes the following remark about perception: 'What a man sees depends both upon what he looks at and also upon what his previous visual-conceptual experience has taught him to see' (Kuhn 1962: 113).
  - The standard examples concern **ambiguous** images (like the duck-rabbit), where prior priming is strongly correlated with the disambiguation opted for by the visual system.

> What happens in perceptual processing, according to this account, is that sensory information is interpreted by reference to the perceivers background theories, the latter serving, in effect, to rule out certain etiologies as implausible causal histories for the present sensory array (Fodor 1984: 31)

Ways of Thinking and Evidence Reports

- As Kuhn and others have noted since (Kuhn 1962), adopting a hypothesis commits one to **seeing the world in its terms**.
- We need a theory **before** we can see – for we do not report what we see as patterns of sensory irradiation.
  - Our observational evidence is reported by the complement clause of *I saw that …* statements we accept – and we don't say 'I saw a brownish quadrilateral region in the left of my visual field' – we say 'I saw a table'.
- When we share our evidence in this way, it is nevertheless dependent on our prior judgments about what there is to be observed – that, e.g., 'table' is a legitimate observational category.
- For a more theoretical example: adopting **Newtonian mechanics** involves rejecting the Aristotelian conception of natural places. After the choice of the new theory, **one no longer sees motion as a matter of things moving by their own nature**.

Consequences

- If theory, and more generally the kinds of conceptual frameworks we bring to bear, influence perception, then what follows?
- For one thing, it looks like the kinds of evidence we'll get ends up **supporting our existing theories**, being partly interpreted already in light of them.
  - So those theories will end up better supported, and it will be correspondingly more difficult to change your mind about those theories. If the language we speak determines a conceptual framework, carrying with it certain principles about the correct taxonomy of reality, then that conceptual framework will be repeated confirmed by further perceptual experience which (not coincidentally, given the foregoing) fits that framework.
- In this sense the frameworks we adopt will make it more difficult to have an experience which doesn't fit the framework, i.e., calls for revision to the framework.

Example: CPI

- Reiss (2013: 150–57) spends a long time explaining how CPI measurement is influenced by a significant array of overt and covert decisions about how and what to measure.
  - What price data is collected, and where;
  - What items have their prices tracked, including whose purchases get counted (Reiss 2013: 156);
  - How to accommodate changing preferences and needs in the population in terms of what they purchase and how much (the 'index number problem') – how to translate from a basket of goods to a fixed quantity of utility to consumers;
  - How to manage changes in quality of items, and how to distinguish such changes from changes in decisions to purchase that are themselves already influenced by changes in price – 'choice cannot be used as an indicator of preference' when old goods become unavailable (Reiss 2013: 155).
- Most relevant to our previous discussion:

> economists tend to be highly theory-driven when they estimate the impact of quality changes on people's well-being. To give just one example, an increased variety of products is always interpreted as a good thing because consumers can satisfy their preferences more precisely. But greater variety is not necessarily a good thing as it increases the costs of decision-making, among other things. What is the true value of having yet another variety of

References

References

Boyd, Nora Mills and James Bogen (2021) 'Theory and Observation in Science', in Edward N Zalta, ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2021/entries/science-theory-observation/.

Cartwright, Nancy (2007) 'Are RCTs the Gold Standard?', *BioSocieties* **2**: 11–20. doi:10.1017/s1745855207005029.

Cartwright, Nancy (2008) 'Evidence-Based Policy: What's to Be Done about Relevance?: For the 2008 Oberlin Philosophy Colloquium', *Philosophical Studies* **143**: 127–36. doi:10.1007/s11098-008-9311-4.

Chemla, Gilles and Christopher A Hennessy (2019) 'Controls, Belief Updating, and Bias in Medical RCTs', *Journal of Economic Theory* **184**: 104929. doi:10.1016/j.jet.2019.07.016.

Douglas, Heather (2000) 'Inductive Risk and Values in Science', *Philosophy of Science* **67**: 559–79. doi:10.1086/392855.

Evans, Imogen, Hazel Thornton, Iain Chalmers, and Paul Glasziou (2011) *Testing Treatments*, 2nd edition. Pinter & Martin.

Fodor, Jerry (1984) 'Observation Reconsidered', *Philosophy of Science* **51**: 23–43.

References (cont.)

Hanson, Norwood Russell (1958) *Patterns of Discovery*. Cambridge University Press.

Hempel, Carl G (1965) 'Science and Human Values', in *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*: 81–96. The Free Press.

Howick, Jeremy, Iain Chalmers, Paul Glasziou, Trish Greenhalgh, Carl Heneghan, *et al.* (2011) 'The Oxford 2011 Levels of Evidence'.

Kuhn, Thomas S (1962) *The Structure of Scientific Revolutions*. University of Chicago Press.

Lewis, David (1973) 'Causation', *Journal of Philosophy* **70**: 556–67. doi:10.2307/2025310.

Mill, John Stuart (1874/1974) *A System of Logic, Ratiocinative and Inductive, Books I–III*, 8th edition, The Collected Works of John Stuart Mill, John M Robson, ed. University of Toronto Press; Routledge & Kegan Paul.

Petrosino, Anthony, Carolyn Turpin-Petrosino, and Sarah Guckenburg (2010) 'Formal System Processing of Juveniles: Effects on Delinquency', *Campbell Systematic Reviews* **6**: 1–88.

Reiss, Julian (2013) *Philosophy of Economics*. Routledge.

Sanderson, Ian (2002) 'Evaluation, Policy Learning and Evidence-based Policy Making', *Public Administration* **80**: 1–22. doi:10.1111/1467-9299.00292.