Research Design in Political Science
POLS4011/POLS8058

Richard Frank
5 May 2026

WEEK 9: WHEN THINGS FALL APART, PART 1 (DIAGNOSING THREATS)

PART 1: OVERVIEW

Weeks 7 and 8 focused on some strategies we use to make causal claims more credible including experiments, natural experiments, instrumental variables, regression discontinuity, difference-in-differences, and process tracing. This week we spend some time talking about things can nevertheless fall apart because even with a well-chosen design, things can go wrong. The data may not measure what you think they measure. The control variables may be doing something different from what you assume. Your sample may be systematically biased in ways your method does not address. The theory you brought to the data may be gradually (and unconsciously) replaced by the patterns you uncover in your data. The job this week is to try and diagnose threats by developing the habit of asking which threats are most damaging, regardless of what methodology you are focused on.

I put this week late in the semester because as you work on research designs the threats we will discuss are not abstract. They are likely to apply directly to your work and to the topics we have covered so far. The framing throughout this week is therefore diagnostic and triage oriented. I want to highlight some challenges you are likely to face as you conduct your research, how you can recognize them in practice, and how to minimize the inferential threats to the greatest extent possible. The goal is not to recite the Shadish et al.(2002) typologies verbatim, but to be able to look at your own draft and say: "these are the three threats that matter most to my inference, here is which is fatal and which is manageable, and here is what I will do about each."

I also have a second motivating reason for focusing on diagnostic causal threats. A fair bit of the published literature can read like a kind of methodological theatre. Authors list the standard threats, claim to have addressed them, and move on. Clarke (2005) is the cleanest demonstration that this routine is sometimes not just unhelpful but actively misleading. Adding more controls does not always reduce omitted variable bias; under specifiable conditions it can actually make the bias worse. Robustness checks performed without thinking about the underlying causal structure are not robustness checks; they are statistical window dressing. The diagnostic approach we are trying to build is a defence against this kind of theatre, in your own work and that of others.

This week also introduces a contemporary set of threats that have less of a settled literature. The data deluge that Anderson (2008) celebrates is now genuinely with us, and the generative AI tools that Bail (2024) surveys have moved from speculation to more or less standard practice. These tools create new opportunities and new threats. The promise is the automation of expensive tasks (e.g., annotation, classification, simulation of survey responses); the threat is that biases in training data, model opacity, and the seduction of plausible-sounding output can amplify rather than fix existing problems of inference. We read both Anderson (2008) and

Bail (2024) because they focus on the same question: what is the role of theory and design once the data become large enough or the tools clever enough that you can find a pattern for almost any claim? The answer in both cases is that theory and design matter more, not less.

Finally, Kalyvas (2004) diagnoses theoretical and logistical threats in a literature I grew up in. He argues that the civil war literature is biased because the practical conditions of data collection (e.g., security, access, language, infrastructure) push researchers toward urban areas and away from the rural areas where most of the violence occurs. No amount of fancy regression models can adequately address this threat. The threat lives upstream of the analysis, in the question of where the evidence comes from and the theoretical perspective many scholars adopt.

Plan for today

- Overview: from designing for credibility to diagnosing threats
- Readings: the canonical typology, omitted variable bias, theory and the data deluge, AI and social science, and the urban bias
- Group activity: the pre-mortem workshop
- Looking ahead to Week 10 (the responses)

Key themes for this week

- Threats to inference are best understood as a diagnostic toolkit, not a checklist to be ticked off.
- The four-fold validity typology (statistical conclusion, internal, construct, external) gives a vocabulary for naming threats precisely.
- Adding more controls does not always reduce bias, and routine "robustness checks" are no substitute for thinking carefully about the data-generating process.
- Theory disciplines empirical work; the data deluge does not eliminate the need for theory, it raises the stakes for getting it right.
- Generative AI offers real potential benefits but introduces new and partly unfamiliar threats to inference that current practice has not yet absorbed.
- Sampling and measurement decisions made early in the research process can produce biases that no later analytical step can fix.

The differentiated expectations continue. Honours students should be able to identify the major threats to inference for their own research design and rank them by severity. MA/PhD students should be able to evaluate threats to inference in published research, distinguish fatal threats from manageable ones, and articulate the design or analytical responses that each calls for.

PART 2: READINGS

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Ch. 1–2. Boston: Houghton Mifflin.

Clarke, Kevin A. 2005. "The Phantom Menace: Omitted Variable Bias in Econometric Research." *Conflict Management and Peace Science* 22(4): 341–352.

Anderson, Chris. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *Wired*, 23 June 2008. https://www.wired.com/2008/06/pb-theory/

Bail, Christopher A. 2024. "Can Generative AI Improve Social Science?" *Proceedings of the National Academy of Sciences* 121(21): e2314021121.

Kalyvas, Stathis N. 2004. "The Urban Bias in Research on Civil Wars." *Security Studies* 13(3): 160-190.

Shadish et al. (2002), Ch. 1 and Ch. 2

Shadish et al. (2002) is the canonical reference for thinking about threats to inference. The book is the third generation of a tradition that runs from Campbell and Stanley (1963) through Cook and Campbell (1979). Each generation refined the typology, but the underlying move is the same: when we cannot achieve experimental control, we should at least be able to name the things that could go wrong, in a vocabulary precise enough that we can argue about which are present and how serious each is.

Chapter 1 lays the conceptual groundwork. The authors define a cause as something whose manipulation would change the outcome, and they emphasise that causation in the social sciences is generally probabilistic and complex (e.g., multiple causes, contingent effects, dependence on context) rather than a more deterministic, single-cause picture that some natural sciences develop. This framing matters because it shapes what we should expect from any given study. We are looking for credible probabilistic claims about average effects under specified conditions, not universal laws. The chapter also distinguishes between molar and molecular causal questions, which is a useful framing: a molar claim treats a complex package ("electoral observation") as the cause; a molecular claim breaks the package into components and asks which is doing the work. Most of our designs estimate molar effects, and this is fine, but it is important to know what kind of claim a given study supports.

Chapter 2 introduces the four-fold typology of validity that gives the threats literature its structure. Statistical conclusion validity asks whether the statistical inferences about covariation are warranted (sample size, the right test, assumptions of the test, multiple comparisons, etc.). Internal validity asks whether the observed covariation between treatment and outcome reflects a causal relationship within the study (selection, history, maturation, instrumentation, regression to the mean, attrition, and so on). Construct validity asks whether the operationalised treatment and outcome correspond to the theoretical constructs they are meant to represent (this connects directly to our Weeks 3 and 4 readings on conceptualisation and measurement). External validity asks whether the causal relationship generalises across populations, settings, treatments, and outcomes (this connects to the case selection and scope material from Weeks 5 and 6).

The threat lists within each category are useful as prompts when reading or designing a study. But they are not meant to be applied mechanically. Shadish et al. (2002) are explicit that the relevant threats depend on the specific design, the specific question, and the specific context. A threat that is fatal in one study may be irrelevant in another. The trick is to develop the discipline of looking at a design, considering the candidate threats, and identifying the few that matter. That is what we will practice in the group activity.

One conceptual point worth flagging is the trade-off the authors describe between internal and external validity. A tightly controlled experiment may have very strong internal validity (the causal claim within the study is credible) but poor external validity (the conditions of the experiment do not match the conditions of the world we want to claim about). The familiar example is the lab-experimental finding that does not replicate in field conditions. The reverse is also true: a study with strong external validity (a representative survey, say) may have weak internal validity (no manipulation of the variable of interest, all the usual selection problems). This trade-off is one we have encountered repeatedly this term, and the four-fold typology gives you the vocabulary to talk about it precisely.

Reading questions

Honours

1. In your own words, explain the difference between internal validity and construct validity. Why does a study with strong internal validity not automatically have strong construct validity?
2. For your own research project, which of the four validity types do you think is most at risk, and why? Identify one specific threat in that category.

MA/PhD

1. Shadish et al. (2002) argue that the relevant threats depend on the design, question, and context. Choose a published study you know well and identify the two or three threats that would be most damaging if they were operative. Are those the threats the authors actually address?
2. The internal-external validity trade-off is often discussed as a structural feature of empirical research. Is it a real trade-off, or is it sometimes an artefact of how studies are designed? Can you think of designs that mitigate it rather than reproduce it?

Clarke (2005)

Clarke's (2005) argument is that the conventional wisdom about omitted variable bias (OVB), which most of us absorb in our first econometrics course (cough, POLS2044, cough) and then hopefully apply, is wrong. We are taught that adding a relevant control variable to a regression reduces the bias on the coefficient of interest. Clarke (2005) shows, with simple algebra, that this is not generally true. Adding a control can leave the bias unchanged, reduce it, or even increase it. The direction depends on the relationships among the included variables, the omitted variable, and the outcome, which is to say, on the data-generating process that we cannot directly observe.

The core result is straightforward. Suppose the true model includes some omitted variable Z that affects Y, and we are trying to estimate the effect of X on Y. We add a control W to the regression. The effect of adding W on the bias of the coefficient on X depends on three things: the relationship between W and X, the relationship between W and Z, and the relationship between Z and Y. Under common configurations of these relationships, adding W increases rather than decreases the bias on the coefficient of interest. Clarke (2005) lays this out with a small set of equations that are accessible if you read them slowly and systematically.

The implications for our research are significant. The standard "robustness check" in political science consists of running the main regression, then running it again with additional controls, and reporting that (ideally) the coefficient on the variable of interest is similar across specifications. The implicit logic is that if the coefficient is robust to adding more controls,

then effect must be real. Clarke (2005) shows that this logic is at best incomplete and at worst misleading. Robustness across specifications can reflect genuine causal effects; it can also reflect an unobserved structure in which the included controls are not actually addressing the bias. Conversely, instability across specifications does not necessarily mean a finding is fragile; it may mean the analyst is changing the bias structure each time they change the controls.

What is the takeaway? Clarke (2005) is not saying we should give up on regression with controls. He is saying we should think harder about what we are doing when we add a control. The right way to address omitted variable bias is to reason explicitly about the causal structure linking the variables, maybe with a directed acyclic graph (DAG), and to add controls that block confounding paths without opening up new ones. This is more demanding than the standard "add more controls" approach, but it is also far more honest about what regression with controls can and cannot accomplish.

There is a pretty clear connection to last week. The IV strategy in Miguel et al. (2004) is, among other things, an attempt to escape the OVB problem by finding a source of variation in the treatment that is plausibly independent of the omitted variables. Clarke (2005) explains why escaping OVB through better controls is harder than it looks, and therefore why design-based strategies like IV (when they work) are valuable. He also explains why we should be sceptical of robustness checks that simply add more controls without justifying why they would address the relevant bias.

Reading questions

Honours

3. In your own words, explain why adding a control variable does not always reduce bias. Use a specific example, real or hypothetical, where you can imagine the bias going up rather than down.
4. If Clarke (2005) is correct that routine robustness checks do not necessarily strengthen a causal claim, what should an applied researcher do instead? What would convince you that a regression-based finding is robust?

MA/PhD

3. Clarke (2005) shows that the bias from adding a control depends on the data-generating process, which is unobservable. How is the analyst supposed to reason about something they cannot observe? What role do theory, prior literature, and qualitative knowledge play in this reasoning?
4. Compare Clarke's (2005) critique of the "add more controls" approach with the design-based strategies we discussed in Week 8. Are these alternative responses to the same problem (OVB), or are they addressing different problems? When would you choose each, and why?

Anderson (2008)

Anderson (2008) is the shortest reading this term and the most provocative I could find. He argues, in essence, that the data revolution has rendered the traditional scientific method we have so painstakingly been working through obsolete. With petabytes of data, we no longer need to formulate hypotheses, build models, or test theories. We can just look for patterns. Correlation, in this view, is enough. "The numbers speak for themselves."

Anderson's (2008) examples are drawn from Google's approach to translation, advertising, and search ranking. None of these systems, he points out, rest on a theory of language, of consumer behaviour, or of relevance. They rest on enormous amounts of data and on algorithms that find patterns in that data. The systems work in the sense that they produce useful outputs. The argument is that this kind of theory-free, pattern-driven approach is the future of all empirical inquiry, including political sciences.

I want to highlight several things that I would be keen on getting your take on in class. First, does the absence of explicit theory mean the absence of theory? Google's systems are theory-laden; the theory is just embedded in the choice of features, the architecture of the model, the loss function, and the data the system is trained on. The theory is implicit, which is not the same as absent. The same will be true of any pattern-finding exercise in the social sciences. Decisions about what to measure, what to compare, and what counts as a meaningful pattern are theoretical decisions whether the analyst notices them or not. Second, are the kinds of questions that pattern-finding can answer a narrow subset of the questions we want answered? I would think that "Which ad will this user click on?" is a different kind of question from "Why did this country democratise?", and the latter requires causal reasoning that pattern-finding cannot supply. Third, does the data deluge solve the fundamental problem of causal inference or does it intensifies it, because spurious correlations are easier to find in larger datasets. Without theory to discipline the search, the data deluge is likely to produce a deluge of false positives.

Anderson (2008) is provocative in some interesting and instructive ways, and versions of his point have emerged in subsequent waves of methodological enthusiasm: machine learning in the 2010s, deep learning in the late 2010s, and LLMs now. Each wave has its share of voices announcing that the new tools have made theoretical reasoning obsolete. One possible response to these arguments is not to ignore the new tools (they have real benefits) but to be clear-headed about what they do and do not solve.

Reading questions

Honours

5. Anderson (2008) claims that with enough data, the scientific method is obsolete. Identify the strongest version of his argument, and then identify what you think is the most compelling counter-argument.
6. If "the numbers speak for themselves," why do we still need theory in political science research? Give a concrete example where theory does work that pattern-finding cannot.

MA/PhD

5. Anderson (2008) describes Google's systems as theory-free. Are they really theory-free, or is the theory just implicit? What follows for the analogous argument about social science?
6. Anderson's (2008) argument has been recycled several times in the past two decades, with new technologies replacing the old as the supposed agent of the obsolescence of theory. What feature of the argument makes it perennially appealing, and what feature makes it perennially wrong?

Bail (2024)

Bail (2024) is the contemporary version of Anderson (2008). He surveys the rapidly developing literature on how generative AI tools, particularly LLMs, are being used in social science research. He neither dismisses the tools as a fad nor accepts the most enthusiastic claims about their transformative potential. The piece is a useful map of where the field is and where its main controversies lie, and for our purposes it is the cleanest statement available of the new threats to inference that LLM-based methods introduce.

Bail (2024) groups LLMs' applications into several categories. First, automated content analysis and classification: LLMs can label text data (sentiment, topic, ideological position) at low cost, replacing or augmenting human coders. Second, simulation of human responses: LLMs can be prompted to simulate the answers a particular type of person would give to a survey, with potential applications to hard-to-reach populations or counterfactual inquiry ("what would moderate Republicans have said about this issue in 1995?"). Third, augmentation of qualitative work: LLMs can summarise interview transcripts, generate hypotheses, and assist in the coding of qualitative material. Fourth, simulation of social processes: agent-based models in which the agents are LLM-driven can be used to study collective dynamics.

Each application carries familiar threats in unfamiliar forms. Automated classification by LLMs raises the question of training data bias: an LLM trained on internet text reflects whatever distribution of views and language is overrepresented in that text, and this bias spreads into any classification it produces. Simulated survey responses raise the question of whether the simulated population matches the real one: there is now a small but growing literature showing that LLM-simulated respondents systematically differ from human respondents in non-obvious ways, often biased toward whatever the model considers the modal or socially desirable response. Augmentation of qualitative work raises the question of whether the LLM is summarising what the interview said or what its training data leads it to expect interviews to say. Agent-based simulations raise the question of whether the agents' behaviour reflects anything we would call human reasoning, or whether it reflects the model's pattern-completion habits.

The threats Bail (2024) lists overlap with older threats (measurement error, sampling bias, construct validity) but it also includes some new ones. The novelty comes from the opacity of the models. We typically do not know exactly what data the model was trained on, what fine-tuning it received, or how it would behave on a hypothetical we have not tested. This opacity makes it hard to characterise the bias structure, which is a precondition for thinking about how to fix it. The result is that the standard tools we have for diagnosing threats (Shadish, Cook, and Campbell's typology, Clarke's warnings about controls) apply, but applying them to LLM-based methods requires additional vigilance because the underlying process is more opaque than the regressions we are used to.

The constructive contribution of Bail (2024) is a set of best practices, including pre-registration of LLM-based analysis, sensitivity tests across model versions and prompts, documentation of training data and fine-tuning where possible, and explicit comparison of LLM outputs against human-coded benchmarks. These are recognisable as analogues of the practices we already use for other empirical methods, adapted to the specifics of generative models. The article is useful for our purposes because it shows that the diagnostic disposition we are trying to build ("what are the threats specific to this design?") extends naturally to new methods, and because it shows that the new methods do not let us escape the threats; they reshape them.

Reading questions

Honours

7. Bail (2024) argues that LLM-based methods carry both familiar threats in unfamiliar forms and novel threats. Identify one of each and explain how they would affect a specific kind of social science research.
8. If you were considering using an LLM to classify survey responses or news articles in your own research, what would you want to know about the model and its training data before trusting the output?

MA/PhD

7. Bail (2024) suggests that LLM-based methods can simulate hard-to-reach populations. What would have to be true about the model for this simulation to be informative, and what would the resulting estimate identify? Is it the same as what an actual survey would identify, or something else?
8. Compare the threats Bail (2024) catalogues for LLM-based methods with the threats Shadish et al. (2002) describe for conventional research designs. Are the validity types still the right framework for organising the new threats, or do the new methods require new categories?

Kalyvas (2004)

Kalyvas (2004) is the applied article for this week, and it is one of the more important methodological articles in conflict studies. The argument is simple to state and devastating in its implications. Civil war scholarship, both quantitative and qualitative, is systematically biased because the practical conditions of data collection (security, access, infrastructure, language, the locations of journalists and NGOs) push researchers toward urban areas. Civil wars, however, are mostly fought in the countryside. The result is that what scholars believe about civil war reflects, disproportionately, the urban experience of civil war, which is in important respects unrepresentative of the broader phenomenon.

The piece is a methodological article in the same sense that the readings on selection bias from Weeks 5 and 6 are methodological articles, but it operates at a different level. Kalyvas (2004) is not pointing at a specific statistical procedure or a specific quasi-experimental design. He is pointing at the upstream process of evidence generation, the stage before any analytical decision is made. He is asking where the information that scholars rely on come from and what systematic distortions does that origin introduce. Once you take the question seriously, a lot of the empirical literature on civil war looks different (at least for me). Claims about ethnic motivations, about ideological conflict, about the role of grievance, all rely on evidence that disproportionately comes from the parts of conflicts where the urban-based information ecosystem reaches. The parts that the information ecosystem does not reach (the countryside, the periphery, the periods when journalists could not travel) are systematically underrepresented.

The threats Kalyvas (2004) identifies are several. The first is sample selection. The cases that make it into the dataset are not a random sample of conflicts; they are the conflicts that produced enough urban-accessible information to be coded. The second is measurement bias within cases. Even within a single conflict, the events that are recorded in datasets are disproportionately the events in cities, which differ systematically from rural events in their participants, motivations, and consequences. The third is theoretical bias. Theories of civil war

that fit the urban data well will be selected for, and theories that would fit rural data better may go undeveloped because the rural data are not visible. This last point is the most serious, because it means the field as a whole develops in a direction shaped by the bias rather than corrected against it.

The connection to Bail (2024) and Anderson (2008) is worth making. Kalyvas (2004) shows that the conditions under which evidence is generated can systematically distort what we think we know, and no amount of analytical sophistication on the back end fixes the problem. Anderson's (2008) celebration of the data deluge ignores this entirely; if the data deluge is itself biased (and most datasets, including the ones LLMs are trained on, are), then more data does not produce more accurate inference. It produces more confidently wrong inference. Bail (2024) is alert to this in his discussion of training data bias, but the deeper Kalyvas-style point applies to any social science dataset, large or small: the conditions of data generation are part of the analysis, not separate from it.

For your own work, the Kalyvas (2004) lesson is to look hard at where your data came from. Who produced them, under what conditions, with what incentives, leaving out what? If you are using a standard dataset (UCDP, V-Dem, Polity, ANES, ESS, World Bank indicators), you are inheriting the biases built into the data generation process. Some of those biases are well documented; many are not.

Reading questions

Honours

9. Kalyvas (2004) argues that civil war scholarship is biased toward urban experience. Identify one substantive claim in the civil war literature that you think might be vulnerable to this bias, and explain how it would be affected.
10. Think about the data sources you intend to use in your own project. Where do those data come from? What kinds of cases or events are likely to be over- or under-represented?

MA/PhD

9. Kalyvas (2004) claims that theoretical development in a field can be shaped by the bias in the available data, with the result that better theories of underrepresented cases never get developed. Evaluate this claim. Is there evidence that civil war theory has actually been distorted in this way, or is the bias mostly affecting empirical claims rather than theoretical ones?
10. Kalyvas (2004) identifies a threat that is upstream of analytical choices. Compare this to the threats discussed in Shadish et al. (2002) and in Clarke (2005). Are upstream threats fundamentally different in kind from downstream threats, or is the distinction one of degree?

Cross-reading question

Across the five readings this week we have seen threats from poor design (Shadish et al.2002), threats from misapplied analytical tools (Clarke 2005), threats from theoretical complacency (Anderson 2008), threats from new technologies (Bail 2024), and threats from the conditions of data generation (Kalyvas 2008). Are these all instances of a common underlying problem, or are they genuinely different kinds of threats? If you could only address to one of them in your own work, which would you prioritise, and why?

PART 3: GROUP ACTIVITY

The Pre-Mortem Workshop

Last week put you in the role of a designer, building a study from a particular methodological starting point. This week I ask a different question: imagine the study you are now writing has been published, has attracted attention, and has been comprehensively dismantled. What killed it? The activity is a pre-mortem in which each group imagines that a project has already failed and works backward to identify what most plausibly caused the failure. This is the opposite to what political scientists often do, where we are usually asked to defend our work, and that is why I think it may be useful. By imagining the failure first, you bypass your defensive reflex and surface the threats you would otherwise rationalise away.

The pedagogical goal is to translate the diagnostic vocabulary in this week's readings into a working habit. By the end of today you should have a short, prioritised list of the threats most damaging to your own design and a concrete plan for what to do about each. This is also the kind of analysis that should appear, in compressed form, in the limitations section of any honest empirical paper, including the research design paper you are writing for this course.

The setup

You will work in small groups of three or four, mixed across Honours, MA, and PhD like last week. Each group will work through every member's design in turn, spending about 8-10 minutes per design.

Before the small-group work begins, take five minutes individually to write a one-paragraph summary of your design. The paragraph should specify, in plain language: the research question, the units of analysis, the treatment or independent variable of interest, the outcome or dependent variable, the research design you are using, and the main data source(s). This paragraph is the object the group will work on. Be specific; vague summaries produce vague pre-mortems.

The pre-mortem (per design)

For each design in the group, work through the following four steps. The presenter reads their paragraph aloud, then sits back and listens. The other group members do most of the talking in steps 2 and 3. Designate one group member as the scribe, rotating across designs so everyone takes a turn.

Step 1: Imagine the failure (1 minute)

As a group, imagine that this study has been published, has been widely discussed, and has been comprehensively criticised. The criticism is now the dominant view. Do not yet say what the criticism is; just stipulate that it has happened.

Step 2: Surface the threats (5 minutes)

List, as quickly as possible, every plausible reason the study failed. Push for quantity over quality; the goal is to surface threats that the presenter would not raise on their own behalf. Use the readings as resources. From Shadish et al. (2002): are there threats to construct, internal, statistical conclusion, or external validity? From Clarke (2005): are the controls actually addressing omitted variable bias, or might they be making it worse? From Anderson (2008) and Bail (2024): is the design using theoretically vacuous pattern-finding, or AI-assisted

methods whose biases are not transparent? From Kalyvas (2004): where do the data come from, and what is systematically left out? Aim for at least eight to ten plausible threats per design.

Step 3: Triage (3 minutes)
Now sort the threats. Put each threat into one of three categories: fatal (if this is true, the study's main claim cannot be sustained), serious (this would weaken the claim significantly but not kill it), or manageable (this can be acknowledged in a limitations section without changing the substantive conclusion). Be honest about which category each threat belongs in. The temptation in real research is to relabel fatal threats as manageable. Resist it here.

Step 4: Repair (1 minute)
For the top one or two threats (those classified as fatal or serious), propose a concrete change to the design that would address them. Not a hand-wave ("we could control for that") but a specific change: a different data source, a different identification strategy, a different sample, a different outcome measure, an additional case study, a falsification test. The presenter writes this down and keeps it.

Reporting back
After the small-group rotations are complete, we will come back together as a class. Each group will share, in two minutes, the most interesting threat they surfaced (across all designs in the group, not your own) and the most surprising repair. We are not naming and shaming individual designs in this part; we are building a shared vocabulary of the threats that turn up most often in the kinds of work the class is doing.

Why this format
I can think of three reasons the pre-mortem may be more productive than a standard "discuss the limitations of your design" exercise. First, the imagined failure removes the defensive frame. You are not being asked to defend your design; you are being asked to predict its critics. This is psychologically easier and produces more honest analysis. Second, the group setting recruits the diagnostic instincts of people who do not have a stake in the design and who are therefore more likely to see threats the presenter cannot. Third, the triage step forces the move from listing threats to ranking them, which is the actual skill I am after this week. A list of twelve threats is methodological theatre; a ranked list with two or three at the top is a research plan.

If you find this format useful, you can run a pre-mortem on your own work at any later stage of the project. The version we are doing today is compressed. A serious pre-mortem on a near-final paper might run an hour with two or three colleagues, and it is some of the most productive time you can spend on a piece of work. The reason researchers usually do not do this is that the academic incentive structure rewards confident assertion of findings, not honest consideration of what could be wrong with them. The skill is learnable, and I think this is a good time to start learning it.

PART 4: WRAPPING UP AND LOOKING AHEAD

Three takeaways from today. First, threats to inference are a diagnostic toolkit, not a checklist. The Shadish et al. (2002) typology gives you the vocabulary, but the application of that vocabulary to a specific design is a judgment call that depends on the question, the data, and the context. The pre-mortem is one way of building the habit of making those judgment calls explicitly. Second, the routine practices of empirical research (adding controls, running robustness checks, reporting standard errors) are not substitutes for thinking about the data-generating process. Clarke (2005) is the cleanest demonstration of this, but the same point applies to Bail's (2024) discussion of LLM-based methods and to Kalyvas's (2004) discussion of upstream sampling bias. The mechanical application of methods is methodological theatre; the diagnostic application of methods is research. Third, theory and design are not made obsolete by data, by AI, or by any other tool. The opposite is closer to the truth. As the tools become more powerful and the data more abundant, the role of theory and design becomes more important, because there are more spurious patterns available for the unwary to find.

Next week is the second half of this final block. We move from diagnosing threats to responding to them. Some threats are addressable through better design (which we covered in the causal inference fortnight). Some are addressable through better measurement (which we covered in Weeks 3 and 4). Some are addressable through transparency, replication, and pre-registration. Some are simply features of your research question and have to be acknowledged honestly in the limitations section. Week 10 will give us a vocabulary and a set of practices for responding to threats, including the open-science practices that have become standard in the past decade and the publication practices that have not yet caught up. Bring your prioritised threat list from today's activity to next week. We will use it.