Evidence for Causal Mechanisms in Social Science: Recommendations from Woodward's Manipulability Theory of Causation

Rosa W. Runhardt*

In this article I analyze process tracing, a causal mechanism-based technique for testing causal claims in the social sciences that requires one to specify a chain of intervening causes between any putative cause and effect. I argue that one should not only give evidence that the intervening causes are present in a suitable case study, as process tracing methodologists recommend, but also provide counterfactual evidence to show that each link in the chain is genuinely causal. I detail what that counterfactual evidence should consist of, using Woodward's manipulability theory, and argue that this evidence relies on tentative comparisons to other case studies.

1. Introduction. In a backlash against the pervasiveness of statistical methods (see King, Keohane, and Verba 1994), in the past decade certain social scientists have focused on finding the causal mechanisms behind observed correlations (see Mahoney 2001; Hedström and Ylikoski 2010; Hall 2013). To provide evidence for such mechanisms, researchers increasingly rely on process tracing, a method that involves contrasting the observable implications of several alternative mechanisms.

In this article I analyze contemporary methodological recommendations for good process tracing (see George and Bennett 2005; Brady and Collier 2010; Bennett and Checkel 2015). In particular, I argue that in order to give evidence for a causal relation between a variable X and Y, not only should one specify the set of intermediate variables, as the methodologists recommend, but one should also provide counterfactual evidence to show that each

*To contact the author, please write to: Department of Philosophy, Logic and Scientific Method, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK; e-mail: r.w.runhardt@lse.ac.uk.

link in the chain is genuinely causal (see Psillos 2004). I detail what that counterfactual evidence should consist of using Jim Woodward's manipulability theory: process tracers must study not only the intervening variables but also the *intervention variables* of each link in the causal chain. After giving the specific requirements for such interventions, I show what complications arise for the social sciences.

This article then is set up as follows. First, I analyze what process tracing is and what it aims to do, giving an example from political science to support my analysis. Second, I set out the relevant aspects of Woodward's theory, and particularly of his notion of an intervention. Third, I evaluate process tracing in light of Woodward's theory and conclude that it indeed lacks evidence for genuine causation. I finish by discussing what an 'interventionist' process tracing would look like.

2. A Philosophical Reconstruction of Process Tracing. Process tracing is a mechanism-based method for analyzing causal relationships. The term refers to two techniques (see Bennett and Checkel 2015), bottom-up and top-down process tracing. Bottom-up process tracing involves surveying a situation of interest with as few preconceptions as possible, in order to then formulate a hypothesis about possible causal connections in that situation. Top-down process tracing tests type-level causal hypotheses about the mechanisms producing the chain of events or 'process' connecting a putative cause and an observed effect, using data collected in case studies. Bottom-up process tracing and top-down process tracing are occasionally mixed; a researcher may start with a bottom-up study to formulate hypotheses and continue with a top-down study to see whether these hypotheses are corroborated or refuted by the evidence available. In what follows, I will look at the second type of process tracing, that is, top-down process tracing, because I wish to evaluate how process tracers justify causal claims.

Let us consider top-down process tracing more formally, before looking at an example from political science. In top-down process tracing, one formulates a hypothesis about what may be the cause of an observed effect, and by what mechanisms the two are connected. One subsequently tries to provide support for one's own hypothesis, as well as refute any existing rival hypotheses, in a case study. In the simplest case (in which there is only one hypothesized mechanism), we may formalize process tracing as follows. Let us call the researcher's own hypothesis H_Z. H_Z holds that a causal mechanism Z is behind a process linking a putative cause, X, and the observed effect, Y. This mechanism has observable implications, that is, a set of variables Z_i such that $X \rightarrow Z_1 \rightarrow Z_2 \rightarrow \ldots \rightarrow Y$ (where $Z_i \rightarrow Z_j$ means that Z_i causes Z_j). It is this chain of events that process tracers trace; note that there is a difference between causal mechanisms and the process that they produce.

The observable implications of the mechanism are generally called 'causal-process observations' (CPOs) in the literature (see Brady and Collier 2010). CPOs can be thought of as the salient observations a process tracer uses to evaluate a causal hypothesis. Bennett and Checkel define the observable implications of mechanisms as "the facts and sequences within a case that should be true if each of the alternative hypothesized explanations of the case is true. Which actors should have known, said, and did what, and when? Who should have interacted with, worried about, or allied with whom?" (2015, 30).

2.1. Case Study.   To illustrate process tracing, I will now analyze political scientist Kirstin Bakke's (2013) study of tactical innovation during the Chechen wars. The tactics of the Chechen insurgents radicalized dramatically between the First and Second Chechen War. The Second Chechen War saw "large-scale hostage-takings, suicide terrorism, and kidnappings" (Bakke 2013, 53), including some infamous large-scale terrorist attacks outside of Chechnya and attacks on civilian targets. Simultaneously, after the first war ended, a large number of transnational insurgents had entered the struggle in Chechnya. Bakke uses process tracing to argue that the two events are causally connected: the transnational fighters contributed to the tactical innovation of local insurgents, and so radicalization did not (just) have an indigenous source. To show that this is the case, Bakke traces the observable implications of two mechanisms by which the transnationals could influence local fighters: relational diffusion and mediated diffusion.

Relational diffusion is a transfer of information through personal contact, for example, interpersonal interactions. In mediated diffusion a third party puts in touch two previously unconnected parties or brokers information between them. According to Bakke, transnational insurgents in Chechnya transmitted "ideas about morally accepted or effective and efficient tactics" (2013, 38) in schools and training camps, and so local insurgents learned and emulated these tactics. Diffusion was especially effective because the transnationals were able to convince locals that the new tactics had been successful elsewhere, and because these tactics did not go against "local norms for acceptable behaviour" (39).

To show that mediated diffusion and relational diffusion were indeed behind the radicalization of tactics in Chechnya, Bakke deduces several observable implications that should be true if these mechanisms were present. The process tracing evidence that Bakke presents for her claim is fourfold:

1. There is evidence for the right 'background conditions' to make diffusion aimed at a radicalization of tactics possible, such as evidence that fighters came to accept the idea of direct attacks against civilians more over time.

2. Before one of the most prominent hostage crises, the Dubrovka/Nord-Ost theater siege, took place, training camps and schools had already been set up. Thus, Bakke claims that "timing-wise, it is plausible that the tactics were a result of learning or emulation via both relational and mediated diffusion" (2013, 56).

3. There is some evidence for the relational and mediated diffusion of tactics in these training camps and schools: for example, evidence of the most prominent transnational insurgent, Emir Khattab, posting videos of suicide bombings online, as well as evidence of Khattab teaching hostage techniques in the training camps. Bakke's evidence of this particular 'intermediate factor' in the causal chain is thin on the ground; she only cites several secondary sources in support.

4. Bakke tries to make her claims more salient with a (counterfactual) remark: "Suicide terrorism, in contrast [to hostage taking], does not have a local historical template among the Chechens, despite centuries of conflict with central rulers. Thus in the absence of outside influence, it is unlikely that the Chechens would have turned to such a tactic" (2013, 58).

Thus, Bakke presents evidence for the observable implications of the diffusion mechanisms, but her evidence for the causal connection between the steps of the process is thin on the ground.

3. Woodward's Manipulability Theory. Now that I have presented my case study, I will show how one might use counterfactual evidence to show that the links of the process chain are genuinely causal, using James Woodward's manipulability theory. I will first turn to the relevant aspects of Woodward's theory, before explicating how we can apply it to process tracing.

Woodward argues that any successful description of a cause–effect relationship must refer to causal factors that can be manipulated to change the phenomenon under study. Specifically, a variable X is a cause of a variable Y if there exists some 'intervention variable' I that we can use to change X, so that X will then, in turn, change Y without any interference of other variables linked to Y. In other words, using I, we can ascertain that X made the change in Y happen.

I have chosen to look at what would happen if the process tracer committed to Woodward's notion of causation, rather than others, for three reasons. Firstly, Woodward's theory provides an alternative to the probabilistic notions of causation that underlie the statistical approaches to which the process tracing methodology was a reaction. Secondly, Woodward's notion is arguably more suited to studying causal mechanisms in social science than the energy-transfer or mark-transmission notions of causation developed for causal mechanisms in areas like physics, and more suitable than mechanist

accounts in terms of entities and activities (see Machamer, Darden, and Craver 2000). Thirdly, Woodward's notion has not been widely applied to areas such as political science and international relations, and therefore this analysis contributes to the literature in philosophy of causation as well as to philosophy of social science.

3.1. Manipulability Theory. The focal point of Woodward's work is his formal set of necessary and sufficient conditions for X to be a (type-level) cause of Y: "a necessary and sufficient condition for X to be a (type-level) **direct cause** of Y with respect to a variable set V is that there be a possible intervention on X that will change Y or the probability of Y when one holds fixed at some value all other variables Z_i in V" (Woodward 2003, 59).

To illustrate the use of the variable set V, consider the following toy scenario: we are interested in a Scandinavian village, asking whether, for its villagers, eating citrus fruit (X) is a direct cause of an absence of scurvy (Y). To answer that question, we cannot just feed the villagers citrus fruit and observe what happens to their health. We need to take into account other variables that may influence this (lack of) scurvy. Hence, we investigate the villagers' diet and find out that they greatly enjoy eating liver; their liver consumption (Z) is very high. What will happen in our experiments to determine the effect of citrus consumption is the following. If we ignore the liver consumption, Z, of the villagers, we will find that no possible intervention on their citrus consumption, X, will change their developing scurvy or not, Y. Simply put, not eating citrus fruit will not mean that the villagers get scurvy. However, if we keep fixed at 0 the variable Z for these villagers, we will find out that there is an intervention on X, that is, making the villagers eat citrus fruit, that will change Y, that is, whether they develop scurvy. We find that if $X = 0$, that is, the villagers do not consume the fruit, then $Y = 1$, that is, they develop the deficiency disease. If they do consume the fruit, that is, $X = 1$, then they do not develop the disease, that is, $Y = 0$.

The notion of a direct cause alone is too basic for a complete theory of causation. Woodward calls our attention to the possibility of a variable X that influences a variable Y along some route but has no total effect on Y because X's influence is always canceled out by other factors (Woodward 2003, 50). To deal with such cases, Woodward introduces the notion of a contributing cause:

> A necessary and sufficient condition for X to be a (type-level) **contributing cause** of Y with respect to variable set V is that
>
> (i) there be a directed path from X to Y such that each link in this path is a direct causal relationship; that is, a set of variables $Z_1, \ldots, Z_n$ such

that X is a direct cause of Z_1, which in turn is a direct cause of Z_2,
which is a direct cause of . . . Z_n, which is a direct cause of Y; and that

(ii) there be some intervention on X that will change Y when all other
variables in V that are not on this path are fixed at some value.

If there is only one path P from X to Y, or if the only alternative path
from X to Y besides P contains no intermediate variables (i.e., is direct),
then X is a contributing cause of Y as long as there is some intervention
on X that will change the value of Y, for some values of the other variables
in V. (Woodward 2003, 59)

As Woodward himself stresses, a direct cause is always a contributing cause,
but a contributing cause is not always a direct cause.

3.2. Interventions.   The notion of an intervention is a crucial part of
Woodward's argument. Let me give an example before presenting Wood-
ward's technical definition.

According to Woodward's theory, introducing a microfinance institution
in a country will be an intervention variable I for investigating whether
taking out microcredit loans (X) causes a reduction in household poverty
(Y) if and only if the following things hold. First, the introduction of the
microfinance institution has to increase the probability that a microcredit
is taken out. Second, there must be no other source of microcredit loans
besides this microfinance institution (so that when we do not introduce the
microfinance institution, no microcredits will be taken out). Third, and this
is more difficult to ascertain in practice, the introduction of the microfinance
institution should not reduce poverty in a way that is unrelated to micro-
credits. If it turns out, for instance, that opening a microsavings account also
reduces households' poverty, and such accounts are offered by the micro-
finance institution, the third demand will fail. We would not be able to tell
whether the microcredit loan or the microsavings account made the dif-
ference. In general, overlooking other ways besides X whereby I may in-
fluence Y clouds our judgement about the relation between X and Y. Fourth
and last, introducing the microfinance institution must be statistically in-
dependent of all variables that reduce poverty by other means than micro-
credit loans. For instance, if we can only introduce the microfinance in-
stitution in regions that have a stable government, this clouds our judgement:
the stability of the government could itself cause an eventual reduction in
households' poverty. Hence, we must ascertain that there are no other ways
in which I can influence Y; if there were, that would mean that I gives us
a misguided picture of the connection between X and Y. (To see the differ-
ence between the third and fourth requirements, consider the following. Both
the third and fourth requirements are violated if there is a factor Z causally

connected to both I and Y but not to X. Requirement 3 only captures cases in which we have $I \rightarrow Z \rightarrow Y$, whereas for requirement 4, the relation between I and Z is unknown. It may, for instance, just as well be that $I \leftarrow Z \rightarrow Y$.)

This brings us to the four requirements in Woodward's definition of an intervention variable for type-level causation:

> I is an **intervention variable** for X with respect to Y if and only if I meets the following conditions:
>
> (1) I causes X.
> (2) I acts as a switch for all the other variables that cause X. That is, certain values of I are such that when I attains those values, X ceases to depend on the values of other variables that cause X and instead depends only on the value taken by I.
> (3) Any directed path from I to Y goes through X. That is, I does not directly cause Y and is not a cause of any causes of Y that are distinct from X except, of course, for those causes of Y, if any, that are built into the $I - X - Y$ connection itself, that is, except for
>
>> (a) any causes of Y that are effects of X (i.e., variables that are causally between X and Y) and
>> (b) any causes of Y that are between I and X and have no effect on Y independently of X.
>
> (4) I is (statistically) independent of any variable Z that causes Y and that is on a directed path that does not go through X. (Woodward 2003, 98)

In short, I is an intervention variable for X with respect to Y when we can use I to check whether X is a (direct or contributing) cause of Y, that is, when we can use I to change X, whereafter X will change Y without interference from other variables causally related to Y. Using I, we will be able to ascertain that X made the change in Y happen. Thus, Woodward makes a distinction between contributing causes X, intervention variables I that we use to analyze whether a variable X is in fact a cause, and intervening variables Z that are the means by which a contributing cause X influences its effect Y.

Woodward claims that one does not need to intervene in practice to support a causal inference. Instead, one could also look for a natural experiment, an intervention that does not involve intentional human action. In the microfinance case, it may well be that there are two regions in the world that are similar in all crucial respects except that one has microfinance in-

stitutions whereas the other does not. If we compared the two, taking into account all the requirements above, and found that in the country without microfinance institutions a larger proportion of households were below the poverty threshold than in the country with microfinance institutions, then this would corroborate the claim that there is a causal relation between taking out microcredits and reduction of the proportion of poor households.

We must make sure that the reason one of the two regions has a microfinance institution while the other does not is not independently affecting the level of household poverty. For example, if the reason a microfinance institution opened in region A rather than region B is that region A has a more stable government, we would expect the stability of the government to be the reason that the regions' household poverty levels differ. Thus, one should study at least two cases, a 'control case' and an 'experimental case', and justify that these two cases are sufficiently similar. I will discuss what we might mean by 'similar' in this context in section 4.

According to Woodward, though it is sufficient to find an actual intervention variable that answers to requirements 1–4, this is not necessary. One does not need to identify an actual intervention variable in order to establish causation; it is sufficient to establish the counterfactual of what would happen if an intervention on putative cause X were to occur. We can evaluate causal inferences based on nonexperimental data by asking whether these inferences "tell us what the result of an appropriate hypothetical experiment would be" (Woodward, forthcoming, 10): "we ask whether the data are such that (in conjunction with appropriate other assumptions) they can be used to infer what the results of the associated hypothetical experiment would be if we were to perform the experiment, although in fact we don't or can't actually perform the experiment." I will now discuss how feasible this is for the social sciences.

4. Process Tracing Evaluated from the Perspective of the Manipulability Theory of Causation. In this section, I will apply Woodward's theory to Bakke's study in order to illustrate how adopting the manipulability theory would change the process tracing technique. In Woodward's framework, we need to show that all links $Z_i \rightarrow Z_j$ of the chain connecting X and Y are genuinely causal, which we can do using one of the three methods outlined above. In all three cases, we need to show that an intervention on each variable Z_i would result in a change in the subsequent variable Z_j.

In contrast, all that the process tracing methodology outlined above requires is that we observe the deductive implications of the mechanism (the CPOs) in a case study, that is, that we observe the intervening variables of the process. To use Bakke's example, it is less satisfying to simply state that schools and training camps were built between the arrival of transnational insurgents and the use of radical tactics than it is to clearly link that it was in

those camps that local insurgents became convinced that using, for example, suicide bombings is an effective and acceptable tactic.

To contrast the two approaches in more detail, consider what further evidence Bakke needs to support the (simplified) claim that 'the presence of transnational insurgents, X, is a contributing cause of local insurgents' increased use of suicide bombings, Y, via the intervening variable of watching videos of suicide bombings, Z'. The manipulationist framework requires her to answer the following:

1. Is X a direct cause of Z? In other words, is there a (human, natural, or hypothetical) intervention on X that will change Y or the probability of Y when one holds fixed all other variables in V at some value?
2. Is Z a direct cause of Y? In other words, is there a (human, natural, or hypothetical) intervention on Y that will change Z or the probability of Z when one holds fixed all other variables in V at some value?

(In practice, this scientist would also investigate the observational implications of alternative mechanisms. I will not discuss this aspect here.)

For the sake of conciseness, I will focus only on question 2 here. Woodward's definition of an intervention variable indicates that in order to answer question 2, we need to know the following:

2*. There exists a variable I_Z, or we can formulate a hypothetical variable I_Z, which
   a) causes Z;
   b) acts as a switch for Z;
   c) does not directly cause Y and does not cause any causes of Y except those on the path $I_Z \rightarrow Z \rightarrow Y$; and
   d) is statistically independent of any variable A not on the path $I_Z \rightarrow Z \rightarrow Y$ that causes Y.

Hence, concretely, what information does Bakke need to gather for finding an intervention variable I_Z, or establishing the counterfactual claim of what would happen if such an intervention on Z were to occur? Let us consider what the requirements for an intervention on Z are. A variable I_Z is an intervention if, firstly, I_Z causes viewing of videos of suicide bombings; secondly, I_Z acts as a switch for the local insurgents' increased use of suicide bombings (i.e., makes whether the insurgents use this tactic independent of any other variables); thirdly, I_Z does not directly or through a path not on $I_Z \rightarrow Z \rightarrow Y$ cause the increased use of suicide bombings; fourthly, I_Z is statistically independent of any variable A not on the path $I_Z \rightarrow Z \rightarrow Y$ that causes the increased use of suicide bombings.

By giving a detailed list of the evidence Woodward requires, we see what is lacking in the process tracing methodology. A process tracer interested in the causal connection between X and Y who follows methodologists such as George, Bennett, and Checkel is not concerned with finding interventions. Rather, they investigate whether the observable implications of all factors (e.g., the presence of transnational insurgents, watching of suicide bombing videos, increase in suicide bombings) are present in some case study (e.g., the Second Chechen War).

Although Bakke carefully collects evidence of the presence of all intervening factors, what she does not do in her work (and what she is not required to do, if we take methodological advice from George, Bennett, and Checkel seriously) is find an actual intervention variable or establish the counterfactual claim of what would happen if an intervention were made. Thus, she does not prove that X, the presence of transnational insurgents, is a contributing cause to Y, the increase of suicide bombings. Finding observable consequences of the three factors does not show that they are causally related, because there may be alternative explanations for these observable consequences. Failure to observe the consequences can falsify, but observing them does not show that the causal hypothesis is correct.

As we have seen, there are several ways in which interventionism can help us give evidence for a causal claim that $X \rightarrow Y$. In the particular context of Bakke's work, an actual human intervention is impossible, and more generally speaking, many social science cases are unlikely to be compatible with this technique.

A natural experiment would rely on finding a sufficiently similar case in which the cause was not present, to see what would happen to the effect. In Bakke's case, we would need to find a (set of) conflict(s) that are similar in every other way to the Second Chechen War, but where transnational insurgents are not present. This requires us to cash out what we mean by 'sufficiently similar'; I will come back to this below.

How about establishing the counterfactual claim of what would happen under an intervention? In that case, Bakke needs to ask, could we have prevented the local insurgents from watching suicide bombing videos, in a way that is in no way connected to their use of this radical tactic through a different route? Would they have used suicide bombings less if we had prevented them from watching such videos? Bakke would have to work out her argument for the earlier-stated counterfactual claim that "suicide terrorism, in contrast [to hostage taking], does not have a local historical template among the Chechens, despite centuries of conflict with central rulers. Thus in the absence of outside influence, it is unlikely that the Chechens would have turned to such a tactic" (Bakke 2013, 58). Although she makes this claim, she does not go into any further detail, and as mentioned above,

Checkel and Bennett never argue that such counterfactual reasoning should be part of the process tracing technique.

The counterfactual claim Bakke makes relies on a similarity comparison, just like a natural experiment would: here Bakke's assumption is that the Chechens before the arrival of the transnational insurgents are sufficiently similar to the Chechens after the arrival of the transnational insurgents, and thus that the diffusion mechanisms that the transnational insurgents set in motion are (one of the) cause(s) of radicalization. Therefore, if there had been no transnational insurgents, we would not have seen the same use of suicide terrorism. A similarity comparison in areas like political science is, however, difficult to defend. Many of the processes that are being traced seemingly occur only once, and there is reason to believe that one can always find causally relevant differences between, for example, two conflicts, or one conflict at different points in time.

5. Conclusion. I have shown that process tracers postulate causal hypotheses that relate a cause X and effect Y by a path of intervening variables $Z_1 \ldots Z_n$. They then find a case study in which both C and E are present and investigate whether $Z_1 \ldots Z_n$ are also present. Woodward defines that X is a *contributing* cause of Y with respect to V if and only if there is a set of intervening variables $Z_1, \ldots, Z_n$ such that X is a direct cause of Z_1, which in turn is a direct cause of Z_2, which is a direct cause of $\ldots Z_n$, which is a direct cause of Y. Process tracing does not establish the complete right-hand side of this "if and only if" statement. Process tracers show that a set of intervening variables exists, but they do not show that each link of the chain is a relation of direct causation.

I have argued that if process tracers were to commit to Woodward's notion of causation, they would have to provide evidence that there is a possible intervention to show that the relations they hypothesize are genuinely causal. I have shown that one way of getting to knowledge of such intervention variables is by comparing and contrasting one case to another, with either a natural or hypothetical experiment, which in both cases requires a sophisticated analysis of to what extent such cases are similar. As such, further work is required to make the process tracing technique useful and feasible.

REFERENCES

Bakke, Kristin M. 2013. "Copying and Learning from Outsiders? Assessing Diffusion from Transnational Insurgents in the Chechen Wars." In *Transnational Dynamics of Civil War*, ed. Jeffrey T. Checkel, 31–62. Cambridge: Cambridge University Press.
Bennett, Andrew, and Jeffrey T. Checkel. 2015. "Process Tracing: from Philosophical Roots to Best Practices." In *Process Tracing: from Metaphor to Analytic Tool*, ed. Andrew Bennett and Jeffrey T. Checkel, 3–38. Cambridge: Cambridge University Press.

Brady, Henry E., and David Collier. 2010. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. 2nd ed. Lanham, MD: Rowman & Littlefield.

George, Alexander L., and Andrew Bennett. 2005. *Case Studies and Theory Development in the Social Sciences*. Cambridge, MA: MIT Press.

Hall, Peter A. 2013. "Symposium: Tracing the Progress of Process Tracing." *European Political Science* 12:20–30.

Hedström, Peter, and Petri Ylikoski. 2010. "Causal Mechanisms in the Social Sciences." *Annual Review of Sociology* 36:49–67.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.

Machamer, Peter, Lindley Darden, and Carl F. Craver. 2000. "Thinking about Mechanisms." *Philosophy of Science* 67 (1): 1–25.

Mahoney, James. 2001. "Beyond Correlational Analysis: Recent Innovations in Theory and Method." *Sociological Forum* 16 (3): 575–93.

Psillos, Stathis. 2004. "A Glimpse of the Secret Connexion: Harmonizing Mechanisms with Counterfactuals." *Perspectives on Science* 12 (3): 288–319.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

———. Forthcoming. "Methodology, Ontology, and Interventionism." *Synthese*. http://link.springer.com/article/10.1007%2Fs11229-014-0479-1.