Transformative Treatments*

L.A. Paul

University of North Carolina, Chapel Hill

University of St Andrews

Kieran Healy

Duke University

Abstract: Contemporary social-scientific research seeks to identify specific causal mechanisms for outcomes of theoretical interest. Experiments that randomize populations to treatment and control conditions are the "gold standard" for causal inference. We identify, describe, and analyze the problem posed by *transformative treatments*. Such treatments radically change treated individuals in a way that creates a mismatch in populations, but this mismatch is not empirically detectable at the level of counterfactual dependence. In such cases, the identification of causal pathways is underdetermined in a previously unrecognized way. Moreover, if the treatment is indeed transformative it breaks the inferential structure of the experimental design. Transformative treatments are not curiosities or "corner cases", but are plausible mechanisms in a large class of events of theoretical interest, particularly ones where deliberate randomization is impractical and

* We are especially indebted to David Danks and Chris Hitchcock for comments and discussion. Thanks are also due to Liam Kofi Bright, Tobias Gerstenberg, Martin Glazier, Paul Horwich, Jenann Ismael, Li Kang, Enoch Lambert, Marc Lange, Christian Loew, Richard Pettigrew, Jonathan Schaffer, Wang Wei, and an anonymous referee.

quasi-experimental designs are sought instead. They cast long-running debates about treatment and selection effects in a new light, and raise new methodological challenges.

An experiment asks the world a question. In social-scientific studies, for instance, researchers often want to know whether there is a causal relationship between some event and an outcome of theoretical interest. Usually, they will also have a story in mind about the nature of the causal mechanism linking the event to the outcome. That is, in addition to inferring *that* the event causes the outcome, researchers often want details of *how* the event causes the outcome. What is the pathway from cause to effect? What is the structure of the causal dependence? How, in detail, does the cause bring about the effect? Such questions concern the metaphysical structure of the *mechanism*: the causal relation and its constituents. Does the cause bring about the effect by kicking off a chain of events leading directly to the effect? Or does it bring it about via preventing other events from happening? Does the cause actively create the effect? Or does it work merely by removing roadblocks, thus allowing the effect to occur unimpeded?

The scientific goal is to identify the specific structure of the causal mechanism at work. For example, if minorities earn a lower average wage than whites, is it because they tend to be lower-skilled, or because employers are choosing not to hire qualified minority workers (Pager, Western, and Bonikowski 2009)? If married people report higher levels of happiness, does that mean marriage makes people happier, or is it that happier people are more likely to get married (Stutzer and Frey 2006)? If friends tend to have similar taste in music, is it because people's social networks influence their tastes, or because people with similar tastes tend to become friends (Vaisey and Lizardo 2010)? The difficulty of observing and isolating causal mechanisms is a core problem for researchers. Experiments are generally seen as the most powerful method for identifying them.

Experiments ask their questions by comparing populations that differ systematically only on some *treatment* of theoretical interest. The goal is to identify the causal mechanism based on differences in the outcomes observed for each population. An ideal study would compare population *P1* in world *W1* with population *P2* in world *W2*, where these populations and worlds are perfect duplicates in all respects until time *t1*. Then, in world *W1*, we simply observe the

population *P1* as it evolves from time *t1* to *t2*. In contrast, in *W2*, at *t1*, *P2* is "treated," that is, an intervention is performed at *t1*.[1] We then observe the population *P2* as it evolves to *t2*. At *t2*, we compare *P2* to *P1* and draw inferences based on this comparison. *W2* is the "treated" world and *P2* is the treated population. *W1* is the "control" world and *P1* is the control population.

Given that the laws and all background conditions in both worlds are the same, and assuming determinism for simplicity, any differences in evolution between *P2* and *P1* can be ascribed to the intervention or treatment. This is an application of the idea that we can discover causation by uncovering counterfactual dependence. *Ceteris paribus*, if we can treat the population *P2*, and we can determine that had the treatment not occurred the effect would not have occurred, we can infer that the treatment causes the effect (Woodward 2003).

Sometimes, the question at issue is merely a matter of simple counterfactual dependence. Our interest, however, is in cases where further information about the mechanism between cause and effect is desired. In these cases we want to know the deeper structure of the causal pathway from the cause to the effect of theoretical interest. This requires researchers to ask just the right question with respect to the theory being tested. Often, the existence of some kind of cause-effect relationship and perhaps also the direction of the empirical association will already be established by prior work, but there will be substantive disagreement about the nature and structure of the causal pathway. If the population-matching process is set up correctly and the correct treatment is applied cleanly, this can allow for the accurate measurement of a generic cause-and-effect relationship, as well as the identification of the specific causal mechanism at work.

Causal Identification

[1] We are not concerned about reductive accounts of modality here, so we can dispense with Lewisian concerns about transition periods, etc. We adopt a version of the approach that ignores the need for a transition period. See Woodward (2003), Maudlin (2007), and Paul and Hall (2013) for further discussion.

Common-or-garden variety causal identification problems arise when we cannot identify the dependence or pathway of interest because of some failure of data, or a shortcoming in the research design. For example, a survey might find a robust correlation between two variables, but be unable to tell us about the direction of counterfactual dependence in that part of the causal structure. Or an experiment might reveal that a treatment reliably brings about an effect, but by a causal pathway that remains opaque. In either case, the results are consistent with multiple, redundant, distinctively different causal stories. Competing causal mechanisms cannot be confidently ruled out.

We might be tempted to say that, at least in social science, we could identify causal mechanisms by using testimony from the individuals involved. Why not just ask the research subjects how the effect of interest was caused, for example with respect to their preferences or their impression of the factors perceived as relevant to their decisions? But testimony is not a reliable guide. Research on questions about causal mechanisms requiring testimony from subjects is complicated not just because different possible pathways could be responsible for the effect, but also because some mechanisms may not be socially desirable to report (as in the case of wage discrimination), may be subject to rationalization or confirmation bias (as in the case of marital happiness), or may be psychologically opaque to the research subject for other reasons. More generally, thinking that testimony can be used to straightforwardly identify causal mechanisms misunderstands its evidential role. As we emphasized above, the goal is to use the experiment to ask the question, and the resulting data to answer it. While subjects' testimony can be an important part of the data gathered and assessed, we cannot identify causal mechanisms of interest by simply asking subjects to tell us what they are. It is the world (via the data and research design) that must answer this question.[2]

Some designs allow for more precision in causal identification than others. Conventionally, studies can be thought of as being either experimental or observational. The difference depends on whether the treatment can be deliberately assigned. In an experiment, a population is selected, and the experimenter randomly (but deliberately) assigns population members to a control group or an experimental group. Members of the control group are not treated, while members of the

[2] See our discussion of the fundamental identification problem below for a related point.

experimental group are. The treatment is administered to the experimental group, and the effects of the treatment are determined by comparison to the control group. The control group functions as the counterfactual comparison class for the individuals in the experimental group.

In an observational study, treatments cannot be randomly assigned. In the social scientific cases we are focusing on, it is either practically impossible, definitely immoral, or simply illegal to randomly assign people to the treatment of interest, such as getting a college degree, having a child, or losing a parent before the age of fifteen. Instead, in effect, groups are observed after the world "assigns" them to the treatment. Researchers use a variety of identification strategies in an effort to make observational studies resemble the experimental ideal as much as possible. For example, they may attempt to observe populations that are matched as closely as possible except for the variable or quasi-treatment of interest (Stuart 2010), or search for natural experiments that can be used to approximate random assignment to a treatment, or construct "instrumental variables" to proxy for endogenous explanatory mechanisms (Angrist and Krueger 1990).

The Fundamental Identification Problem

Even when they are well-designed, both observational and experimental studies are limited by the *fundamental identification problem* (Holland 2006; Rubin 1974, 2005).[3] The problem comes from the fact that we cannot simultaneously observe the same *individual subject* in a treated and an untreated state. Put another way, we cannot observe the same subject over actual and counterfactual states. This problem does not arise from errors in research design or weaknesses in data collection. Instead, it stems from constraints on the detectability of facts about individuals in empirically inaccessible modal states. We do not have a "modalscope" that lets us observe

[3] A note on terminology: the sort of identification involved in the fundamental identification problem is probably best described as a kind of "mathematical" identification. Some identification problems are better described as "mathematical" if, for example, they involve the mathematical problem of establishing the values of structural parameters in a model. Other identification problems are better described as "causal" if they concern our ability to make an inference from the research design and its results about the structure of the causal relations being uncovered by the science. Our focus here is on causal identification, but causal identification and mathematical identification can be theoretically intertwined when a theory is expressed as a structural model of some kind.

individuals in non-actual worlds. (And at the level of individuals, we do not have subjects who are perfect duplicates, such that the untreated individual can serve as the control for the treated individual.)

Observational studies use various clever methods to manage the problem so as to approximate the so-called "gold standard" of randomized experiments. But even in the experimental case, counterfactual effects cannot be observed at the level of individual subjects. Again, this is because our research is restricted to the actual world:

> "The researcher wants to compare the outcomes that people would experience if they were to receive alternate treatments. However, treatments are mutually exclusive. At most, the researcher can observe the outcome that each person experiences under the treatment that this person actually receives. The researcher cannot observe the outcomes that people would have experienced under other treatments. These other outcomes are counterfactual. Hence *data on treatments and outcomes cannot by themselves reveal treatment effects*" (Manski 2003, 15, italics added).

Why is this a problem? Again, one might say, why can't we just ask individual subjects how they would have responded if they had not been treated? Or, for data that does not involve testimony, why can't we just assess the counterfactual properties of untreated individuals by intelligent projection, describing the way the world would have evolved forward if there had been no intervention, and using this as a basis for assessing the counterfactual? The answer is apparent from the last part of Manski's remarks. The scientist wants *the world* to tell her the answer to her question. If we must determine the counterfactual truths by asking subjects for their *post hoc* assessments, or by relying on *a priori* judgments about how the world would unfold, or by imposing some other theoretical interpretation on the data to establish which counterfactual worlds are closest, then data on treatments and outcomes cannot by themselves reveal which causal mechanism is at work.

To summarize: the scientist is using the experiment to ask the world a question she does not know the answer to. If she must rely on her own theoretical interpretation or the opinionated

interpretations of her subjects to supply the answer, her experiment has not done its job. As a result, the empirical inaccessibility of counterfactual worlds raises a serious problem, since she cannot use empirical methods to make the comparison she needs in order to answer the question. (Recall that it is impossible, even in principle, to observe what happens in nonactual worlds.) In this way, the fundamental identification problem stems from empirical constraints on our access to deep metaphysical facts.

In response, in order to allow the world to provide an answer, scientists change the question. Instead of asking a question about an individual, they ask a question about a group. They replace the estimation of individual treatment effects with average treatment effects.[4] Focusing on the counterfactual problem of observing both treated and untreated states, scholars have developed sophisticated frameworks to estimate average treatment effects across groups of subjects given what we can observe in the actual world (Rubin 2005, Pearl 2009). Researchers attempt to create comparison classes for causal inferences using similar populations, such that, despite internal heterogeneity at the level of the individual, the populations may be considered to be duplicates (with respect to the properties of interest), *at the group level*. The comparison classes for causal inferences use properties of groups rather than properties of individuals and their nonactual, duplicate counterparts, and report average effects rather than individual effects.

Our discussion of the fundamental identification problem is intended to draw attention to the importance of the idea that empirical data on treatments and outcomes is what reveals treatment effects, not *post hoc* rationalization or *a priori* speculation. When metaphysical constraints prevent data on treatments and outcomes from revealing treatment effects, as they do with the fundamental identification problem, methodological revisions and restraint in corresponding interpretations are in order.

---

[4] Because we do not have duplicate individuals in the actual world, we cannot determine individual-level treatment effects by assigning one individual to treatment and using her duplicate as the control. We randomize to create what for practical purposes are actual-world duplicate groups (matched with respect to the properties of interest). The matching is at the level of groups, not at the level of individuals. Then we calculate the average outcome for the treated actual world group and the average outcome for its untreated, actual world counterpart. The difference between these two values is the *average treatment effect*.

So far, our discussion has been uncontroversially within the scope of current thinking on causal inference from experimental and observational data. If, however, we shift our attention to a related question concerning *whether we observe the same subject over time*, a new type of metaphysical difficulty arises. This new problem stems from empirical constraints on the detectability of certain persistence facts about individuals and their preferences across temporal and modal contexts.

Radical preference change

Becoming a Vampire

Consider the following thought experiment:

> *Vampires:* In the 21st century, vampires begin to populate North America. Psychologists decide to study the implications this could have for the human population. They put out a call for undergraduates to participate in a randomized controlled experiment, and recruit a local vampire with scientific interests. After securing the necessary permissions,[5] they randomize and divide their population of undergraduates into a control group and a treatment group. At *t1*, members of each group are given standard psychological assessments measuring their preferences about vampires in general and about becoming a vampire in particular. Then members of the experimental group are bitten by the lab vampire.
>
> Members of both groups are left to go about their daily lives for a period of time. At *t2*, they are assessed. Members of the control population do not report any difference in their preferences at *t2*. All members of the treated population, on the other hand, report living richer lives, enjoying rewarding new sensory experiences, and having a new sense of meaning at *t2*. As a result, they now uniformly report very strong pro-vampire preferences. (Some members of the treatment group also expressed pro-vampire

[5] We acknowledge that securing permission from the IRB is perhaps the most fantastical aspect of this thought experiment.

> preferences before the experiment, but these were a distinct minority.) In exit interviews, all treated subjects also testify that they have no desire to return to their previous condition.

Should our psychologists conclude that being bitten by a vampire somehow satisfies people's underlying, previously unrecognized, preferences to become vampires? No. They should conclude that being bitten by a vampire causes you to *become a vampire* (and thus, to prefer being one). Being bitten by a vampire and then being satisfied with the result does not satisfy or reveal your underlying preference to be a vampire. Being bitten by a vampire *transforms* you: it changes your preferences in a deep and fundamental way, by replacing your underlying human preferences with vampire preferences, no matter what your previous preferences were. In this example, the physical nature of the treated subjects is changed in a way that replaces their original psychological natures with new ones.

This thought experiment illustrates a case where there is no principled empirical way to distinguish between a treatment that causes one's underlying preferences to be revealed (or, alternatively, to be concealed), and a treatment that causes one's preferences to be *replaced*. In a case like this, there is no principled way to use data on treatments and outcomes to distinguish between a causal mechanism that reveals existing preferences (which are then satisfied) and a causal mechanism that replaces existing preferences with new preferences (which are then satisfied). This raises a causal identification problem: we cannot identify the specific causal mechanism at work. It might be objected that we know the subjects have had their preferences replaced rather than revealed, because preference replacement is what happens when you get bitten by a vampire. But in our thought experiment, the pathway to the changed nature of the treated subjects is evident only by construction. We are assuming that we already know the physiologically transformative consequences of a vampire bite on people's psychological preferences about vampirism. In general this will not be the case—ordinarily, we are using experiments to try to discover this sort of fact.

The source of the problem involves subtle metaphysical assumptions embedded in experimental design. An implicit assumption of experimental studies is that the members of the treatment

group (and the control group) have stable, persisting selves over time. Colloquially, we assume that the members of the treatment group are the same individuals or the same people over time, and in particular that they are the same individuals before and after the experiment. That is what allows us to infer that there has been an effect on an individual by comparing her testimony at *t1* with her testimony at *t2*.

In such cases, to detect a meaningful change in a subject, we assume that the individual members of the population have stable, persisting selves over time, reflected in the assumption that the individuals' relevant underlying preferences (to be human, to be a vampire) remain stable. Given the assumption of stability, we can take ourselves to be discovering the nature of the preferences of a subject across a state change by attending to the results of the experiment.

However, in our vampire study, these stability assumptions fail. In this example, the trouble is that new preferences, including the relevant underlying preferences, are created by the treatment and replace the original preferences. If stability of self is tracked by stability of underlying preferences, the treatment (being bitten by a vampire) effectively creates new selves, and so is effectively creating a new population after *t1*. The new population is a population of vampires, and vampires have preferences that are very different from those of humans.[6]

It is obvious that an experiment that physically replaced its treatment population with a new group of people at the moment the treatment was applied would be evidentially worthless. Imagine, if you like, that the subjects are led into the Treatment Room. Then a research assistant enters and quickly pushes them all out a side door, and a new "post-treatment" group is ushered in to take their place. In our vampire study, the replacement is also obvious. The original human selves have been replaced with vampire selves.

Of course, the individuals participating in our vampire study are changed physically, and may even count as new persons. However, the persistence question we wish to attend to concerns

[6] In metaphysical terms, we might hold that these individuals are strictly personally identical over time (their origins are the same, their spatiotemporal trajectory is continuous, and their bodily properties are largely the same). However, even if the criteria for personal identity can be met, the psychological selves have not persisted.

psychological persistence, not personal identity. We use the language of preferences here because that is the established terminology, especially in the literature on assessing treatment and selection effects in the social sciences. But to be clear, by "underlying preferences" we mean what could be thought of as the core elements of one's character or basic personality traits, etc: those aspects that, were they to be replaced, would lead us to say that someone's identity-defining values had been replaced. Independently of whether the criteria for personal identity are met, in our vampire study, the psychological selves in the experimental population have been replaced. And it is the psychological selves, realized by the relevant cluster of underlying preferences, that are of experimental interest.[7]

Vampire treatments are not presently being reported in the social science literature. However, a large class of real world events may involve psychological transformations of a relevantly similar kind in physically persisting individuals. Our thought experiment highlights the assumption that may not hold: that the relevant underlying preferences of individuals under study persist through treatment. It is easy to see what justifies this assumption. We have more or less the same people, or individuals, at *t1* as we have at *t2*. However, what we cannot assume is that this sort of personal continuity ensures the relevant sort of psychological continuity.

Thinking like an Economist

Next, consider the following case:

> *Economists*: In the late 20th century, economists begin to populate North America. Psychologists decide to study the implications this could have for the human population. They put out a call for undergraduates to participate in a randomized controlled experiment, and recruit a local economics professor with scientific interests. After securing the necessary permissions, they randomize and divide their population of undergraduates into a control group and a treatment group. At *t1*, members of each group are given standard assessments for selfishness and altruism. Then members of the experimental group are taught Introductory Economics by the lab economist.

[7] See Paul (2014), especially the discussion of finkish preferences in the Afterword

> Members of both groups are left to go about their daily lives for a specified period of time. At *t2*, they are assessed. Members of the control population do not report any difference in their views about selfishness and altruism at *t2*. A very large number of the treated population, on the other hand, report increased levels of selfishness and decreased rates of altruism at *t2*. In exit interviews, many of these subjects say that, in retrospect, their pre-treatment lives were filled with irrationality and poorly thought-out decisions, and testify that they would never return to their previous condition. (Some members of the treatment group also expressed fully pro-selfishness preferences before the experiment, but these were a distinct minority.) Some subjects testify that taking the course changed the kind of person they are. Others say it helped them see their decision-making in a new way.

Three things make this case a useful intermediate one for the purposes of our argument. First, while vampires are imaginary monsters confined to thought experiments, economists actually exist. The case follows a real and intermittently experimental line of research on the possibly causal association between studying economics and becoming more selfish (Marwell and Ames 1981, Carter and Irons 1991, Frank et al. 1993, Bauman and Rose 2011). Second, while the treatment is suspected to have strong effects on preferences, it is still within the reach of actual randomized controlled trials, unlike some of the more intensive real-life cases that we shall discuss below. Observational studies in the literature—i.e., those starting after the point where students have chosen their courses of study—cannot rule out the possibility that Economics students are simply antecedently more selfish than those who take other courses, and have "selected into" the discipline on that basis. But an experimental design can fix this problem. In our case, the experimenters can assign students to the treatment and control groups. This establishes the required counterfactual dependence, making it possible to calculate the average treatment effect on selfishness of taking an Economics class.

Finally, the case also highlights the deeper question about which particular causal pathway is at work. Having established a treatment effect of taking a course, we want to know what the mechanism is. Some scholars working on this topic argue that taking an economics course is a

kind of indoctrination: it *replaces* the preferences of the students who take it. A warm-hearted freshman enters the class and is curdled into a selfish sophomore. An alternative interpretation is more pedagogically generous. Taking the class is an educational experience where students are given the tools to better understand the rational choices they have always wanted to make. On this view, the treatment simply educates students by giving them the means to *reveal* or express their preferences.

Under the first interpretation, the treatment directly creates a preference to be more selfish. Under the second interpretation, the treatment reveals an underlying preference to be more selfish. According to the latter, let us say, the treatment removes something that is inhibiting or preventing the manifestation of the disposition. If a treatment removes the preventer of a disposition to be more selfish, we can think of the causal pathway as an instance of what is sometimes called "double prevention" (Paul and Hall 2013, 215–231).

The experiment cannot adjudicate between these alternative pathways. Each pathway demonstrates the same counterfactual dependence. Under each interpretation, the effect (that the individual becomes more selfish), counterfactually depends on the application of the treatment. In both interpretations, the treatment is what causes the effect of being more selfish, but in the first case the treatment's effect is brought about directly, while in the second it is brought about via double prevention. The pathways are substantively different, but from the experimenter's point of view they are indistinguishable.

Testimony from treated students will not settle the question. Observers might worry that those subjects who insist the class helped them become more rational people have in fact been indoctrinated or brainwashed. Conversely, we might be skeptical of the students who claim to have had their preferences transformed. Perhaps they are now simply a little more educated and unwilling to own up to their earlier foolishness.

Can we investigate further, in hopes of experimentally distinguishing between the pathways? We could make some progress by very carefully measuring selfishness before and after treatment. For example, we might stratify or "block" the experimental design on the basis of pre-test scores

on some carefully-calibrated selfishness scale. Subjects who pre-tested as highly selfish would then be taken as a subgroup and randomly split into treatment and control conditions. Subjects who pre-tested as highly *un*selfish could be another subgroup, and those in the middle of the scale yet another. Random assignments within these subgroups would allow us to calculate more fine-grained average treatment effects, conditioning on pre-existing levels of selfishness. This would be important if, for example, we thought that taking economics only has an effect on students of average selfishness, with already very selfish or very unselfish students unaffected.

Refining our randomization strategy in this or similar ways will further refine our estimate of the treatment effect and its scope conditions, but will not adjudicate between the causal pathways of interest. An experiment can assign the treatment of taking the course, and this allows us to calculate an average treatment effect for it. But to clarify the mechanism, what would be needed in addition is some means of "treating on *P*" at the level of the individual subject, where *P* is the capacity to have one's preferences drawn out (that is, revealed), *versus* created or brought into existence, as a result of being exposed to the treatment itself. We do not, at present, have any means of assigning subjects in this way.

Of course, if we are already committed in advance to one or other of the candidate mechanisms—to the view that the treatment reveals preferences, or to the view that it creates them—then we will be tempted to take the results as evidence in favor of our preferred pathway. A significant treatment effect in the expected direction is certainly better evidence for our theory than finding no observed difference between the treatment and the control groups. But the competing theories remain underdetermined with respect to the data. *Both* predict an effect in the same direction, just by different causal pathways. The difference between the competing pathways is substantively consequential. Our assessment of a classroom where preference changes were brought about by ideological replacement would be quite different from one where existing views are drawn out or nurtured by the educational process. An individual whom we believe to have been indoctrinated would be viewed quite differently from one we believe to be better able to express their own views as the result of a good education.

The underdetermination in the *Economists* case illustrates, in a small way, the possibility of a much larger and more widespread phenomenon. There is a significant class of ordinary life "treatments" of both general and social-scientific interest where persisting individuals may undergo psychological transformation such that their psychological selves or relevant underlying preferences could be replaced, rather than merely revealed. Such events may include, but are not limited to, having a child; participating in intense military combat; suffering massive physical trauma; undergoing intense training in medical or law school; or having a sudden religious conversion. People who have gone through these experiences often report that their beliefs and preferences are radically different now in ways that make their previous lives seem alien to them, or that they now have core beliefs or values that their previous selves would have rejected.[8] We can call interventions that replace preferences in this way *transformative treatments*.

Transformative Treatments

As already noted, in the social sciences most treatments cannot be randomly assigned. Our evidence about their effects is almost always observational. Moreover, many of the most theoretically interesting questions seem to implicate these sort of unassignable and potentially transformative treatments. Social scientists wonder what they would discover if they really *could* randomly assign people to treatments like *having a child* or *being in combat* or *converting to Christianity*. For many treatments of interest, random assignment would be a big improvement over *post hoc* controls, instruments, or matching. And so, in practice, researchers spend a lot of time trying to make their observational data approximate the experimental ideal. In the past two decades, there has been a strong move in the social sciences towards implementing randomized controlled trials as widely as possible. The reason for this is clearly stated by Angrist and Kreuger (1999): "... the prevailing view is that the best evidence about counterfactuals is

[8] They may tell us that the transformative event "made them into a new person". In our terms, this should be understood as the claim that their psychological selves have changed, usually dramatically. Colloquially, such individuals may describe themselves as "not the same person as before", but this simply reflects the ambiguity between "same self" and "same person" in ordinary expressions. At the very least, we do not take them to imply the denial of strict personal identity through this change. It's not as though having a child or going to war implies that you need to get a new driver's license.

generated by randomized trials, because randomization ensures that outcomes in the control group really do capture the counterfactual for a treatment group."

Causal identification strategies in the social sciences attempt to empirically find and statistically take advantage of exogenous or quasi-exogenous variation on a variable of interest, and make that play the role of randomized assignment to treatment and control groups, so that the data can more closely or plausibly approximate the standard of a randomized trial. Meanwhile, efforts to carry out actual experiments are also very much on the rise. This may tempt us to think that *the* central limitation on social-scientific knowledge is our practical inability to enforce random assignment to treatment and control conditions, and that any move closer to the experimental standard improves our knowledge of the causal mechanisms we are investigating. However, the challenge posed by transformative treatments cannot be dealt with through randomization. This is because it arises from the treatments themselves.

With a transformative treatment, the assignment procedure itself creates new (psychological) populations. If a treatment creates a new population, a central advantage of randomization is lost. That is, we lose the ability to "ensure that outcomes in the control group really do capture the counterfactual for a treatment group". When the new population is created from the experimental group after *t1*, the experimental group is no longer the population that was drawn from the original, randomized population constructed for the study. Thus, at *t2*, the control group drawn from the original randomized population can no longer function as an informative counterfactual comparison, at least in the intended sense, to the treatment population at *t2*. So even if we were able to suspend the rule of law and randomize at will, we would not get ideally informative experiments. Instead, we would have a machine that creates populations with new preference profiles.

Transformative treatments introduce a kind of incoherence into the causal model, a principled mismatch of populations, and one that stems from complications involving the metaphysics of counterfactual dependence. Recall our discussion of the fundamental identification problem. Because we cannot evaluate the treatment effect by measuring the evolution of treated individuals against that of their (nonactual) untreated counterparts, we must create a control

group using our actual population. So, our starting pool of subjects is randomly divided into a treatment group and a control group. Both groups are allowed to evolve forward through time, one treated, one untreated, to time *t2* for assessment. In effect we are simulating, within the actual world, the evolution of an actual treated world (the treatment group) paired with its counterfactual untreated world (the control group). Ordinarily, at *t2* the treated group is compared to its control counterpart in order to assess relevant changes and to determine the causal mechanism. And remember: the treatment group and the control group are counterparts at the level of *groups*, so our results must be taken as average effects rather than individual level effects.

If a treatment is transformative, at *t1*, the treatment—in addition to introducing the properties whose causal effects we wish to track—causes some core, self-defining preferences of the subjects in the treated group to be replaced. The possibility of transformative treatments creates two problems. The first is one of underdetermination. The changes we observe in the treated group relative to the control group are consistent with more than one causal mechanism—in our cases, with one mechanism that reveals a pre-existing preference and with another that creates a new preference. But since the problem is introduced by the treatment itself, we cannot refine the control group prior to treatment in order to eliminate the underdetermination. The second problem that is created lies with the control group. In these cases, if the treatment is indeed transformative, by imposing the treatment we effectively destroy our ability to use the control group to informatively track relevant changes in the treatment group. The underdetermination permits the possibility that the preference has been transformed. And if the preference *has* been transformed, the experimental group no longer matches the control group in the relevant sense.

To restate the second problem: we want to make a comparison between the control group and the experimental group at *t2*, but the randomization is only applied to the experimental group that existed at *t1*, before the treatment. The control group, *C,* at *t1*, can function as a counterfactual match for the experimental group at *t1*. But if the treatment is transformative, *C* is rendered ineligible as a control at *t2*. That is, at t2, *C* is no longer a counterpart, under the intended similarity relation, for the experimental group. (For a different problem of mismatch between

causal variable and effect variable, see Halpern and Hitchcock 2010. For related work on how changes in the world can require changes in models, see Kummerfeld and Danks 2014.)

Why not eliminate the possibility of underdetermination by just asking individual subjects whether their preferences have been revealed or whether they have been replaced? As with the fundamental identification problem, we cannot solve the problem this way. Recall the purpose of the experimental approach: the scientist wants the world to tell her the answer, and wants to use the experiment to provide data on treatments and outcomes that can identify the causal mechanism. If we must identify the causal mechanism by asking subjects for their *post hoc* assessments, data on treatments and outcomes cannot by themselves reveal which causal mechanism is in play.

This does not mean that we cannot include this sort of testimony as part of a broader assessment of our data. Subjects' testimony about their preferences being changed or tranformatively replaced remains informative. It can alert the researcher to the possibility that her control group from *t1* can no longer function as a control at *t2*. But this evidence cannot rule out competing mechanisms, and so it cannot demonstrate that the change in preferences really is the result of transformative replacement. The experiment is what demonstrates the causal pathway. Relying on the testimony of the agents is not sufficient to demonstrate the underlying mechanism of interest, and so the mechanism remains underdetermined.

The problem with underdetermination is most evident when researchers are interested in uncovering mechanisms that participants cannot or will not give reliable testimony about, or in getting evidence about something the study population may not tell us directly in a reliable way. Consider our opening examples involving disputes over the causal mechanism that underlies wage gaps, happiness in marriage, or tastes in friendship. Or consider studies of implicit bias, where direct testimony from participants confirming or denying their bias is not necessary or even sufficient to identify the causal mechanism at work.

Furthermore, candidate causal mechanisms proposed by these studies typically differ substantively in the interpretation they give to the role played by preferences. The implications

for interpretation and further intervention will be very different depending on whether a treatment is taken to have replaced the relevant preferences of subjects or simply to have revealed preferences they had all along. These are not even the only alternatives. Perhaps a post-treatment change in testimony reflects adaptive preference formation—i.e., subjects have reconciled themselves to their new state. Or we might wonder whether subjects even have the capacity to recognize that their preferences were so adapted. Possibilities multiply. Could subjects be engaging in a massive act of regret-aversion? Or rationalization? Or avoiding cognitive dissonance? These questions arise even if subjects testify that their preferences have in fact been transformed. In such cases, and in any case that admits of the possibility of transformation, we need a principled (empirical) method for determining whether subjects had previously unrecognized preferences that were revealed on treatment, are presently engaged in a treatment-induced act of rationalization about their new state, or have undergone a treatment that was genuinely transformative.

The *Vampires* example illustrates the conceptual possibility of a treatment that in effect replaces the pre-treatment population with a new population. The *Economists* example showed how even fine-grained randomization to treatment and control groups might not distinguish between causal pathways of interest when those pathways involve the possibility of preference change that is activated by the treatment. Testimony from study participants *post hoc*, meanwhile, can alert us to the possibility that our treatment was transformative, but will not settle the matter. In the *Economists* example, while the scope of the hypothesized replacement of preferences was more modest, it was also quite real and in principle within the scope of experimental studies that could actually be carried out.

To reiterate: there are two separate problems, and each one carries over to the kinds of cases we really care about—for example, those we began with, regarding the effects of childbirth, marriage, job discrimination, and so on. In these high-stakes cases, the transformative potential of the treatment is *prime facie* plausible, true random assignment is typically impossible, and the second-best solution involves post-treatment methods meant to approximate the experimental ideal. In these high-stakes cases, the possibility of underdetermination of the causal mechanism opens the door to the problem, and the possibility that the actual causal mechanism is

transformative breaks the inferential structure of the quasi-experiment. Thus, empirical work in these kinds of high-stakes cases may not in fact be informative about causal processes in the way that researchers often take it to be.

Elucidating the conceptual structure of transformative treatments and identifying the possibility of underdetermination shows that there is a need for methods that can distinguish, in some principled way, between the possibility that preferences are revealed, and the possibility that preferences are transformed. Beyond that, even if we could *determine* that a particular treatment was transformative, such a determination remains grist for our mill. If, in the cases of interest, treatments are in fact transformative, experimental designs to test the effect of such treatments instead create a series of observational studies on distinct populations. Thus, the research design is not doing what we took it to be doing. The destructive effect of transformative treatments on the inferences we can draw using observational data means that we face a second methodological challenge. If transformative treatments exist, new techniques will be needed to study them and new interpretations of the data will need to be considered.

Conclusion

Heckman (2006, 327) remarks that social scientists must grapple with the inconvenient fact that, in a non-ideal experimental context, causal inferences require assumptions. One such assumption is that relevant preferences of experimental populations are not replaced by experimental treatments. We have sought to provide a framework for understanding the causal structure of inferences involving treatments where this assumption does not hold.

Theory in social science is often most in dispute just at the point where individual preferences are underdetermined by observational data. Some theories identify preference-driven selection processes as the cause behind most outcomes. Others emphasize external pressures or social forces. There is also evidence of sour grapes, regret-avoidance, and dissonance reduction in *post hoc* testimony. These disputes are particularly pointed in the case of major events that are consequential for people's lives, from the effects of professional socialization, to induction into the Army, incarceration, family formation, migration, and beyond. Unable to deliberately

randomize assignment to these treatments, social scientists look for populations that approximate this ideal, in order to partially capture the benefits of a true experimental design.

Our argument implies that when treatments are plausibly transformative, the move to quasi-experimental designs is less informative than commonly thought. Transformative treatments "reach into" a subject and significantly change her preference profile at the point of intervention. In effect, across all individuals, the transformative treatment creates a new population. The immediate difficulty is that the possibility of a transformative treatment allows the data to underdetermine the causal pathway. The foundational problem is that if the treatment is indeed transformative then it breaks the inferential structure of the experimental design. Both these difficulties carry over to the interpretation of the results. Researchers may illegitimately favor the revealed preference pathway, and the causal leverage they want will not be available. The inferential advantage of quasi-experiments is thus weaker than typically imagined for transformative cases, even though this is just where observational researchers now aspire to the experimental approach.

We interpret our conclusion as a methodological challenge. To draw a comparison, recognizing the metaphysical inaccessibility of nonactual individual counterparts in the Fundamental Identification Problem did not put an end to empirical research involving random assignment to treatment and control conditions. Nor did it suspend efforts to learn from data or to find ways to approximate individual-level treatment effects. But it did demonstrate the existence of a basic metaphysical fact constraining certain sorts of inferences from experimental data, and it prompted the development of further methods to generate estimates of average causal effects. We should look for similar innovations to face the present problem.

References

Angrist, Joshua D. and Alan B. Krueger**.** 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics.* 106:979–1014.

Bauman, Yoram and Elaina Rose. 2011. "Selection or indoctrination: Why do Economics Students Donate Less than the Rest?" *Journal of Economic Behavior & Organization* 79:318–327.

Carter, John R., and Michael D. Irons. 1991. "Are Economists Different, and If So, Why?" *Journal of Economic Perspectives*, 5: 171–177.

Correll, Shelley J., Stephen Benard, and In Paik. 2007. "Getting a Job: Is there a Motherhood Penalty?" *American Journal of Sociology* 112: 1297–1338.

Frank, Robert H., Thomas Gilovich, and Dennis T. Regan. 1993. "Does Studying Economics Inhibit Cooperation?" *Journal of Economic Perspectives* 7: 159–171.

Lewis, David. 1997. "Finkish Dispositions." *Philosophical* Quarterly 47:143-158.

Manski, Charles. 2003. "Identification Problems in the Social Sciences and Everyday Life." *Southern Economic Journal* 70:11-21.

Heckman, James. 2006. In William Breit and Barry T. Hirsch (eds.), *Lives of the Laureates: Eighteen Nobel Economists*, fourth ed. New Delhi: Academic Foundation, pp.299–334.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–960.

Kummerfeld, Erich, and David Danks. 2014. "Model change and methodological virtues in scientific inference." *Synthese* 191:2673–2693.

Marwell, Gerald and Ruth Ames. 1981. "Economists Free Ride, Does Anyone Else?" *Journal of Public Economics* 15: 295-310.

Maudlin, Tim. 2007. “A Modest Proposal Concerning Laws, Counterfactuals, and Explanation.” In Tim Maudlin, *The Metaphysics Within Physics*, Oxford: Oxford University Press.

Pager, Devah, Bruce Western, and Bart Bonikowski. 2009. “Discrimination in a Low-Wage Labor Market: A Field Experiment.” *American Sociological Review* 74: 777–799.

Paul, L.A. and Ned Hall. 2013. *Causation: A User’s Guide.* Oxford: Oxford University Press.

Paul, L.A. 2014. *Transformative Experience.* Oxford: Oxford University Press.

Rubin, Donald. 1974. “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.” *Journal of Educational Psychology* 66:688–701.

Rubin, Donald. 2005. “Causal Inference Using Potential Outcomes.” *Journal of the American Statistical Association* 100: 322–331.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*. Second Edition. New York: Cambridge University Press.

Stuart, Elizabeth A. 2010. “Matching Methods for Causal Inference: A Review and a Look Forward.” *Statistical Science* 25:1-21.

Stutzer, Alois, and Bruno S. Frey. 2006. “Does Marriage Make People Happy, or Do Happy People Get Married?” *Journal of Socio-Economics* 35: 326–347.

Vaisey, Stephen, and Omar Lizardo. 2010. “Can Cultural Worldviews Influence Network Composition?” *Social Forces* 88 (4), 1595–1618.

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