Causal Attribution Bounds: Decomposing the Effects of Multiple Causes

Link to latest version

Jason Weitze[1]

November 7, 2025

Abstract

When faced with multiple causes, researchers often ask, "How much did each cause contribute to their combined effect?" For instance, advertisers want to *attribute* the effect of an ad campaign to each ad, while economists want to *decompose* the effect of a policy bundle. This seemingly simple question hides a fundamental problem: there is no single, agreed-upon estimand, so competing methods naturally yield contradictory results. This paper confronts this challenge by developing a novel framework for causal attribution. I characterize a causal decomposition, establishing that it attributes to each cause its individual effect plus some convex share of its jointly-produced effects (e.g., interaction or indirect effects). The result is not a single causal decomposition, but rather a set, formally capturing the fundamental ambiguity of attribution. While standard practice masks this ambiguity by reporting a single point estimate, I propose embracing it by reporting a new estimand: attribution bounds. These bounds transparently communicate the minimum and maximum attributable to a cause across all causal decompositions, collapsing to a single point when there are no joint effects. To make the bounds practical, I provide design-based and observational methods for identification and estimation. Finally, I demonstrate the bounds utility in two applications: one where wide bounds underscore the limitations of reporting a single point estimate, and another where tight bounds indicate minimal ambiguity.

[1]jweitze@stanford.edu, Stanford University and Stanford Data Science. I thank Guido Imbens, Jann Spiess, and Alessandra Voena for their invaluable mentorship and advice. I also thank Michael Baiocchi, Lea Bottmer, Matthew Brown, David Bruns-Smith, Jacob Conway, Tom Cunningham, Han Hong, Mariana Guido, Ramesh Johari, Bing Liu, Lorenzo Rigon, Rose Tan, Amar Venugopal, and Justin Young for helpful comments and discussions.

1. Introduction

From public policy to marketing, researchers often ask not just if a set of causes has an effect, but "How much does each cause contribute to their combined effect?" The allure of this question lies in its promise to distill a complex causal system into a single number per cause, whether *attributing* the effect of an ad campaign to each ad or *decomposing* a policy bundle's effect to isolate the impact of its components (e.g., Econsultancy and Google Analytics 2012).

However, the attractive simplicity of this question hides a fundamental problem: a wide range of competing methods, from sequential decompositions to the Shapley value, naturally lead to contradictory answers (Shapley 1952). The conflict arises not from statistical uncertainty, but from conceptual ambiguity: there is no single, agreed-upon estimand, leaving researchers to make an ad-hoc choice. The prevailing practice is to select one method and report its point estimate as the definitive answer. This, however, masks the underlying problem, presenting one possible answer as if it were the only one.

This paper confronts this challenge by proposing a novel framework for causal attribution. I begin by characterizing a *causal* decomposition, establishing that it must attribute to each cause its individual effect, while allowing for any split of its jointly-produced effects (e.g., an interaction or indirect effect). This characterization reveals that the ambiguity in attribution is fundamental, stemming from the different ways one could attribute the joint effects. Instead of masking this ambiguity, I propose embracing it. To do so, I introduce a new estimand, attribution bounds, corresponding to the minimum and maximum attributable to a cause across all causal decompositions. These bounds transparently communicate the degree of ambiguity, collapsing to a single point only when there is none (i.e., when there are no joint effects). To make this framework practical, I develop identification and estimation strategies for the attribution bounds and demonstrate their value in two empirical applications.

To ground the analysis, I consider a simple yet general setting: decomposing the combined effect of two sequential, binary causes. The sequential nature of the causes clarifies the causal structure by defining a clear direction of influence, avoiding any confusion that may arise when causes occur simultaneously. This paper focuses on decomposing the causes' combined effect: the difference between the realized outcome and the counterfactual had neither cause occurred. The challenge is that not all decompositions of the causes' combined effect are equally principled: a naive method

can assign full credit to a cause that had no effect. The goal, therefore, is to distinguish such methods from principled ones by formally characterizing the set of causal decompositions.

The characterization of causal decompositions proceeds in two complementary parts. First, I take a constructive approach, building from a simple algebraic identity that partitions the combined effect into its constituent components. This partition reveals a crucial distinction between effects that are individually produced by a single cause (direct effects) and those that are jointly produced by multiple causes (e.g., indirect and interaction effects). While individual effects are straightforward to attribute to a single cause, jointly produced effects create a fundamental ambiguity. For instance, in an indirect effect, does the first cause deserve credit for initiating the chain of events or the second cause for its ultimate impact? I argue that both causes have a reasonable claim to the joint effects, leading to the following definition: a causal decomposition attributes to each cause its individual effect, while allowing any split of the joint effects.

Second, I provide a formal argument that independently leads to the same characterization. I propose four properties that a decomposition should satisfy to be considered causal. These properties are designed to be intuitive; for example, a "No Credit for No Effect" property ensures that a cause with no possible effect on the outcome receives no credit. The main theoretical result shows the equivalence of these two approaches: a decomposition is causal if and only if it satisfies the four proposed properties.

This characterization of a causal decomposition provides a common language for classifying and comparing existing methods. First, it establishes a clear boundary between principled and unprincipled approaches, thereby ruling out common heuristics like "last-touch" attribution used in marketing (e.g., Econsultancy and Google Analytics 2012; Quantcast 2012). Second, it reveals that principled methods, including various sequential decompositions, are simply different ways of attributing jointly produced effects. Even the Shapley value, often celebrated for its axiomatic uniqueness, does not uniquely dictate how to allocate credit for the joint effects. Indeed, a recent proliferation of Shapley variants has emerged by adapting game-theoretic ideas to address causality in distinct ways (e.g., Dalessandro, Perlich, Stitelman and Provost 2012; Frye, Rowat and Feige 2020; Heskes, Sijben, Bucur and Claassen 2020; Singal, Michailidis and Ng 2021; Budhathoki, Minorics, Blöbaum and Janzing 2022; Singal, Besbes, Desir, Goyal and Iyengar 2022; Jung, Kasiviswanathan, Tian, Janzing, Blöbaum and Bareinboim 2022). This divergence underscores the fundamental ambiguity in attribution that this paper makes explicit.

The existence of a set of causal decompositions motivates my main proposal: instead of reporting a single number, I advocate for reporting attribution bounds. These bounds indicate the full range of values a cause can be assigned under any causal decomposition, thereby making the degree of ambiguity explicit. To be clear, attribution bounds differ from traditional identification bounds. While identification bounds communicate the limitations of data for pinning down a given estimand, attribution bounds capture ambiguity in the choice of estimand. The width of the bounds thus serves as a direct measure of this conceptual ambiguity, widening with the magnitude of jointly-produced effects and collapsing to a single point in their absence. Reporting attribution bounds replaces a single, potentially misleading estimate with a more transparent and informative summary of a cause's contribution.

I then develop strategies for identifying and estimating attribution bounds in both experimental and observational settings. In an experimental context, the analysis reveals a surprising challenge: no standard experiment can, by itself, identify the attribution bounds. I show that identification is possible, however, with a mild assumption on the sign of the average joint effects and an eligibility experiment that randomizes eligibility to receive treatment rather than randomizing treatment assignment. This design is essential, as directly randomizing treatment assignments overrides individuals' endogenous choices that generate the indirect effects one aims to attribute. In observational settings where experiments are infeasible, I rely on a sequential unconfoundedness assumption (Robins and Greenland 1992; Imai, Keele and Yamamoto 2010b), which generalizes the familiar selection-on-observables condition to the case of multiple, sequential causes.

To demonstrate the practical utility of attribution bounds, I estimate them in two contrasting applications. First, I examine a well-known framing experiment where the jointly produced effects turn out to be substantial (Brader, Valentino and Suhay 2008). Here, the estimated bounds are wide, revealing a high degree of ambiguity that a single point-estimate would obscure. This result highlights the danger of relying on a single method, which could lead to misleading conclusions or allow for strategic reporting. Second, I re-analyze a factorial experiment on early-childhood interventions where the design mechanically eliminates most joint effects (Gertler, Heckman, Pinto, Zanolini, Vermeersch, Walker, Chang and Grantham-McGregor 2014). In this setting, the estimated attribution bounds are tight, revealing that the ambiguity in attribution is empirically negligible. Together, these results, illustrate how attribution bounds function in practice: they transparently quantify the ambiguity inherent in a problem,

providing a more honest summary of a cause's contribution than any single number.

Finally, I distinguish between two distinct applications of attribution: decision-making and communication. While decision-makers often use attribution to guide choices, like allocating an ad budget, this is often misguided. If a decision problem can be fully specified and its solution implemented, doing so will yield a better outcome than relying only on attribution. Furthermore, ambiguity over attribution does not imply ambiguity in decision-making; a single optimal decision can exist even when attribution bounds are wide. The primary role of attribution, therefore, is not direct decision-making. Instead, I argue attribution bounds are ideal for conveying the causal structure of a system in a simple, transparent way; for example, when the ultimate decision problem is not yet known, when there are frictions preventing the implementation of a complex optimal policy, or when presenting results to diverse audiences (Andrews and Shapiro 2021). This leads to a concrete recommendation: optimize decisions directly when you can, and use attribution bounds for clear communication when you cannot.

This paper contributes to the extensive literature on multiple causes, building on foundational work in mediation analysis and sequential treatments (e.g., Robins and Greenland 1992; Pearl 2009; Imai, Keele and Tingley 2010a; Hernán and Robins 2020). Unlike prior work that decomposes the effect of a single cause in the presence of potential mediators (e.g., VanderWeele 2014; VanderWeele and Tchetgen Tchetgen 2014), I focus on decomposing the combined effect of multiple causes. My work also builds on recent critiques of regression in multi-cause settings (e.g., Goldsmith-Pinkham, Hull and Kolesár 2024; Bugni, Canay and McBride 2025), by demonstrating that common decompositions of the causes' combined effect have similar, undesirable properties. I then characterize the entire set of estimands that satisfy key desirable properties and, as a novel reporting strategy, propose reporting bounds on the set itself instead of making an ad-hoc choice from within it.

This work also builds a bridge between the "Effects-of-Causes" (EoC) literature, which measures causal effects, and the "Causes-of-Effects" (CoE) literature, which seeks to explain outcomes or effects (e.g., Mill 2011 [1843]; Gelman 2011; Gelman and Imbens 2013). The question we consider, "How much did each cause contribute to their combined effect?" has the flavor of a CoE question. Typically, CoE questions focus on explaining a single, realized outcome (e.g., "Did the patient recover because of a new drug?"), where the core challenge is the missing counterfactual for a single unit, making even the one-cause case difficult (e.g., Greenland and Robins 1988; Pearl 1999; Rosenbaum 2001; Yamamoto 2012). Instead, my work focuses on explaining an EoC, an average

causal effect. This shift in focus makes the single-cause case relatively straightforward but reveals new challenges in the multi-cause case: there's a fundamental ambiguity over how to attribute jointly-produced effects.

Moreover, this paper is situated within marketing science's extensive work on multi-touch attribution (e.g., Kitts, Wei, Au, Powter and Burdick 2010; Shao and Li 2011; Dalessandro, Perlich, Stitelman and Provost 2012; Li and Kannan 2014; Li, Kannan, Viswanathan and Pani 2016; Du, Zhong, Nair, Cui and Shou 2019; Singal, Besbes, Desir, Goyal and Iyengar 2022). This literature has primarily focused on finding a single best attribution method. In contrast, I argue that the search for a single, generic answer masks an important underlying ambiguity. While some propose circumventing this ambiguity by formally incorporating a specific downstream decision problem, this is often not feasible (e.g., Li, Kannan, Viswanathan and Pani 2016; Berman 2018; Kelly, Vaver and Koehler 2018). For these common instances, I provide a formal framework that shifts the goal from selecting a single method to transparently reporting the range of plausible answers via attribution bounds.

Finally, this work offers a new perspective on the path-dependency problem in decompositions (e.g. Fortin, Lemieux and Firpo 2011). The economics literature on decompositions is epitomized by the simple Oaxaca–Blinder decomposition, which assumes away any ambiguity in attribution (e.g., Kitagawa 1955; Oaxaca 1973; Blinder 1973). More generally, however, a well-known challenge is that results are said to be path-dependent as they depend on the choice of decomposition (e.g., DiNardo, Fortin and Lemieux 1996; Fairlie 2005; Machado and Mata 2005; Bauer and Sinning 2008; Chernozhukov, Fernández-Val and Melly 2013). In this paper's terminology, each decomposition takes a different stance on how to allocate credit for jointly-produced effects. Some researchers respond by averaging over all paths, as with a Shapley value (Shorrocks 2013), while others report results from several paths as a robustness check (e.g., Badinski, Finkelstein, Gentzkow and Hull 2023). My work provides a formal foundation for this last strategy, replacing ad hoc robustness checks with attribution bounds that capture the full range of causal answers. This reframes the multiplicity of reasonable decompositions not as a methodological nuisance to be averaged away, but as a fundamental, quantifiable feature of causal attribution.

2. Conceptual Framework

This section develops the framework for attribution presenting two complementary arguments to define and characterize the set of causal decompositions. The analysis demonstrates that a fundamental ambiguity arises whenever multiple causes jointly produce an outcome. To isolate this conceptual challenge from the identification problem at the core of causal inference (Holland 1986), this section consider an idealized setting where all counterfactuals are known, and yet the ambiguity persists. I return to the more realistic scenario with unobserved counterfactuals in Section 3.

After establishing the notation and defining the causes' combined effect (Section 2.1–2.2), I characterize a causal decomposition in two ways. The first is a constructive argument (Section 2.3), building the definition of a causal decomposition from the structure of the combined effect. The second is a formal argument (Section 2.4), demonstrating that this same definition is characterized by four desirable properties.

The characterization highlights that the ambiguity in attribution is fundamental. This finding motivates the central proposal of the paper, reporting attribution bounds, which are the subject of the following section.

2.1. Setting: Two Sequential Causes

Here, I introduce the notation, setting, and a simple running example.

Consider an individual i subject to two sequential causes represented by the variables, A_{i1} and A_{i2}. Each cause takes on a specific binary treatment value: $\{0, 1\}$. For instance, the first cause could be an ad featuring Colin Kaepernick; the treatment $A_{i1} = 1$ indicates an individual saw the ad, while $A_{i1} = 0$ indicates they did not. Similarly, the second cause could be an ad featuring Serena Williams. Thus, while there are only two causes, there are four treatment sequences, $A_i = (A_{i1}, A_{i2})$, that an individual can be exposed to:

$$A_i = (A_{i1}, A_{i2}) \in \mathbb{A} \equiv \Big\{(0,0), (1,0), (0,1), (1,1)\Big\}.$$

After both periods, an outcome $Y_i \in \mathbb{Y} \subset \mathbb{R}$ is observed. This outcome could be binary, such as a purchase decision ($\mathbb{Y} = \{0, 1\}$), or continuous, like profit ($\mathbb{Y} = \mathbb{R}$).

To formalize the causal relationships, I use the potential outcomes framework. Each individual's outcomes are a function of the entire treatment sequence they receive. An individual's potential outcome function, $Y_i(\cdot) \in \mathcal{Y} = \{y(\cdot) : \mathbb{A} \to \mathbb{Y}\}$, maps each hypothetical treatment sequence to a counterfactual outcome. Equivalently, each individual

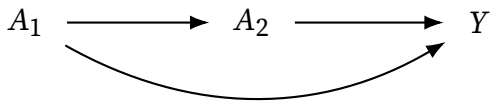

FIGURE 1. A causal Directed Acyclic Graph (DAG) with two sequential causes and one outcome. The first cause can influence the outcome and the second cause. The second cause in turn can also influence the outcome.

has four potential outcomes, $Y_i(0,0)$, $Y_i(0,1)$, $Y_i(1,0)$, and $Y_i(1,1)$, one for each possible treatment sequence. The observed outcome, Y_i, is the potential outcome corresponding to the realized treatment sequence: $Y_i = Y_i(A_{i1}, A_{i2})$.[1]

A fundamental feature of this sequential setting is that the first cause can affect the second cause (see Figure 1). For example, seeing the Kaepernick ad, might lead someone to see the Williams ad or conversely, prevent them from seeing it. I model this dynamic by introducing a potential treatment function for the second cause, $A_{i2}(\cdot) \in \mathcal{A}_2 = \{a_2(\cdot) : \mathbb{A}_1 \to \mathbb{A}_2\}$. This function is analogous to the potential outcome function: it describes the second-period treatment individual i would receive as a function of their first period treatment. Equivalently, each individual has a pair of potential treatments, $(A_{i2}(0), A_{i2}(1))$.[2] As with outcomes, the observed treatment is the potential treatment corresponding to the realized first treatment, $A_{i2} = A_{i2}(A_{i1})$.

Each pair of potential treatments, $(A_{i2}(0), A_{i2}(1))$, characterizes a treatment response type akin to a compliance type in the Local Average Treatment Effects (LATE) framework (Imbens and Angrist 1994). In this analogy, the first cause, A_{i1}, serves as the instrument for the second cause, A_{i2}. For instance, an individual with $(A_{i2}(0), A_{i2}(1)) = (1,1)$ is an always-taker, as they receive the second treatment regardless of the first treatment, while an individual with $(A_{i2}(0), A_{i2}(1)) = (0,0)$ is a never-taker, as they never receive the second treatment. Similarly, an individual with increasing potential treatment, $(A_{i2}(0), A_{i2}(1)) = (0,1)$, is a complier as they only receive the second treatment if they receive the first treatment. Conversely, defiers, $(A_{i2}(0), A_{i2}(1)) = (1,0)$, have a decreasing response, as receiving the first treatment prevents them from receiving the second treatment.

These components—initial treatment, potential treatment function, and potential outcome function—jointly determine an individual's realized outcome and all relevant

[1] This is the standard consistency assumption, which I apply here to outcomes and in the next paragraph to the second-period treatment. The assumption is formalized in Section 3 when discussing identification and estimation.

[2] The mediation literature refers to potential treatments as 'potential mediator values' (e.g., Imai, Keele and Tingley 2010a).

counterfactuals. Thus, this tuple describes an individual's type.

DEFINITION 1. *A **type** θ is a tuple $(a_1, a_2(\cdot), y(\cdot))$, consisting of an initial treatment, $a_1 \in \mathcal{A}_1 \equiv \mathbb{A}_1$, a potential treatment function, $a_2(\cdot) \in \mathcal{A}_2$, and a potential outcome function, $y(\cdot) \in \mathcal{Y}$. We denote the space of types $\Theta = \mathcal{A}_1 \times \mathcal{A}_2 \times \mathcal{Y}$, with typical element $\theta \in \Theta$.*

I adopt lower-case notation to describe an arbitrary type to distinguish the abstract type from a specific individual's realized values (like A_{i1} or Y_i). In general, I write $a_2(a_1)$ to highlight the second treatment's potential to depend on the first treatment, but on occasion, I use a_2 as shorthand for the realized value $a_2(a_1)$.

2.2. The Combined Effect and Its Decompositions

This subsection defines the causes' combined effect and the general concept of a decomposition function. To motivate the fundamental problem of attribution, I contrast the ambiguity of multi-cause decompositions with the clarity of a single-cause scenario. Using several examples of decomposition functions as illustrative examples, I work to build intuition and emphasize a crucial point: decompositions are not inherently causal. The following subsection then characterizes the subset of causal decomposition functions, whose outputs this paper calls attribution values.

The first step is to define combined effect, the quantity that this paper seeks to decompose.

DEFINITION 2. *For $\theta = (a_1, a_2(\cdot), y(\cdot)) \in \Theta$, the causes' **combined effect** is the difference between the potential outcome of the realized treatment sequence and of no treatments:*

$$\tau(\theta) = y(a_1, a_2(a_1)) - y(0, 0)$$

This effect captures the full impact of the treatment path for causal type θ ($y(a_1, a_2(a_1))$), relative to a counterfactual where neither treatment occurs ($y(0, 0)$). For example, a marketing analyst seeking to explain the increase in revenue resulting from an advertising campaign with two ads, is interested in explaining in the combined effect, $\tau(\theta)$.[3]

The goal is to attribute the causes' combined effect to the two constituent causes based on the role they played in creating it. First, we define a generic decomposition

[3]This definition of a combined effect is conceptually distinct from other causal effects of interest, but they can often be seen as special cases. For instance, the effect of a fixed treatment bundle, $(x_1, x_2) \in \mathbb{A}$, is $y(x_1, x_2) - y(0, 0)$. This corresponds to the combined effect $\tau(\tilde{\theta})$ for type $\tilde{\theta} = (\tilde{a}_1, \tilde{a}_2(\cdot), y(\cdot))$ with initial treatment $\tilde{a}_1 = x_1$ and constant potential treatments $\tilde{a}_2(\cdot) = x_2$.

function as a function that splits the combined effect into two components, one for each cause.

DEFINITION 3. *A **decomposition** is a vector-valued function $\psi : \Theta \to \mathbb{R}^2$, that maps a type, $\theta \in \Theta$, to a vector, $(\psi_1(\theta), \psi_2(\theta))$, that sums to the corresponding combined effect:*

$$\psi_1(\theta) + \psi_2(\theta) = \tau(\theta)$$

By convention, ψ_t is the component corresponding to the t^{th} cause for $t = 1, 2$. The space of all such decomposition functions is denoted Ψ.

This definition is primarily mathematical; its only causal feature is that the components must sum to the combined effect, $\tau(\theta)$. Beyond that, it permits arbitrary rules. Consider for instance, Last Touch Attribution, a widely used approach in advertising (Econsultancy and Google Analytics 2012; Quantcast 2012).

EXAMPLE 1 (Last Touch Attribution). *Last Touch Attribution gives the entire combined effect to the cause that provided the last treatment: $\psi^{LT}(\theta) = \big(a_1[1-a_2]\,,\ a_2\big)\tau(\theta)$. In our advertising example, $\psi^{LT}(\theta) = (0, \tau(\theta))$ whenever one sees the second ad ($a_2 = 1$), while $\psi^{LT}(\theta) = (\tau(\theta), 0)$ if they only see the first ad ($a_1 = 1, a_2 = 0$). This method is appealing because of its simplicity, but it does not assign credit based on each causes' effects. For instance, even if the first ad is solely responsible for the outcome (e.g., $y(x_1, x_2) = x_1$), the second ad will receive credit whenever one sees it.*

To apply causal language to decompositions like Last Touch Attribution would be misleading: even though the combined effect has a causal interpretation, the credit assigned to each cause is entirely divorced from the underlying causal structure. This highlights the need for a principled distinction between arbitrary decompositions and those that are causal. The following subsections provide this distinction by formally characterizing the set of causal decompositions, which formally excludes heuristic methods like Last Touch Attribution.

A more principled approach is to use a sequential decomposition, which attributes the combined effect by "turning off" the causes one by one, attributing the change in potential outcomes at each step to the manipulated cause. This approach immediately leads to two important sources of variation. First, in what order should the causes be turned off? Second, how should the other causes be handled during the manipulation: by exogenously holding them fixed or by letting them endogenously respond to the change? The following examples illustrate how these choices lead to different, equally

principled decompositions. I begin with exogenous sequential decompositions, which hold other treatments fixed during manipulation.

EXAMPLE 2.1 (Exogenous Sequential Decompositions). *An exogenous sequential decomposition turns off one cause at a time while holding the other treatment fixed; this yields two decompositions depending on the order of manipulation. The first decomposition,* $\psi^{exogSD,1}$*, corresponds to turning off the first cause and then the second. It first attributes to the first cause, the effect of its removal while holding the second fixed, and then attributes the remaining effect to the second cause:*

$$\psi^{exogSD,1}(\theta) = \big(y(a_1, a_2) - y(0, a_2)\,,\;\; y(0, a_2) - y(0, 0) \big)$$

The second decomposition, $\psi^{exogSD,2}$*, reverses this by first turning off the second cause and then the first. It first attributes to the second cause, the effect of its removal while holding the first fixed, and then attributes the remaining effect to the first cause:*

$$\psi^{exogSD,2}(\theta) = \big(y(a_1, 0) - y(0, 0)\,,\;\; y(a_1, a_2) - y(a_1, 0) \big)$$

Unlike Last Touch Attribution (Example 1), both decompositions assign each cause credit based on a well-defined manipulation of the respective cause.

The exogenous approach assigns each cause credit based on a manipulation that holds other causes fixed. An alternative is the endogenous sequential decomposition, which assigns credit based on a different manipulation: the effect of turning off a cause, while allowing other causes to respond endogenously.

EXAMPLE 2.2 (Endogenous Sequential Decompositions). *An endogenous sequential decomposition turns off one cause at a time, allowing the other cause's treatment to respond endogenously; for instance, turning off the first cause, can change the second treatment from* $a_2(a_1)$ *to* $a_2(0)$*. This yields two additional decompositions depending on the order of manipulation. The first,* $\psi^{endogSD,1}$*, turns off the first cause and then the second. It attributes to the first cause the effect of its removal, including any downstream effects on the second cause, and then attributes the remaining effect to the second cause:*

$$\psi^{endogSD,1}(\theta) = \big(y(a_1, a_2(a_1)) - y(0, a_2(0))\,,\;\; y(0, a_2(0)) - y(0, 0) \big)$$

The second decomposition, $\psi^{endogSD,2}$*, reverses this by turning off the second cause and then the first. Because of the causes' sequential nature, an intervention on the second cause cannot*

affect the first; here, allowing the first cause to respond endogenously is the same as holding it fixed. The decomposition first attributes to the second cause the effect of its removal, and then attributes the remaining effect to the first:

$$\psi^{endogSD,2}(\theta) = (\, y(a_1, 0) - y(0, 0)\,,\;\; y(a_1, a_2(a_1)) - y(a_1, 0))$$

As with the exogenous sequential decompositions (Example 2.1), both endogenous sequential decompositions assign each cause credit based on a well-defined manipulation of the respective cause.

Intuitively, each of these sequential decomposition appears to be causal in the sense that each cause gets credit for the effect of a well-defined manipulation. However, they will generally disagree: the sequential decompositions coincide only when there are no interaction effects or indirect effects.[4]

This multiplicity of causally meaningful decompositions presents a challenge: which one should a researcher choose? One prominent solution is based on the Shapley value, which resolves the ordering problem by averaging (Shapley 1952; Heskes et al. 2020).

EXAMPLE 3 (Causal Shapley Values). *Originating from cooperative game theory, the Shapley value offers a specific solution to the ordering problem by treating each ordering symmetrically. It assigns credit by averaging a cause's contributions across all possible orderings of a given family of sequential decomposition. The difference between these two variants depends on the set of sequential decompositions being averaged (see Example 2.1 and 2.2):*

$$\psi^{exogSV} = \frac{1}{2}\left(\psi^{exogSD,1}(\theta) + \psi^{exogSD,2}(\theta)\right)$$

$$\psi^{endogSV} = \frac{1}{2}\left(\psi^{endogSD,1}(\theta) + \psi^{endogSD,2}(\theta)\right)$$

I refer to the resulting Shapley value variants as an exogenous and endogenous Shapley value, respectively.

Despite its appeal, the Shapley value is not a silver bullet: averaging exogenous sequential decompositions will, in general, give a different answer than averaging endogenous sequential decompositions. Indeed, Heskes et al. (2020) note that what this paper calls the exogenous and endogenous Shapley values coincide only if there are no indirect effects.

[4] We define these effects more carefully in the next subsection.

Moreover, the Shapley value's symmetric average is just one possible summary. In fact, any weighted average across the family of sequential decompositions can constitute a meaningful summary, potentially mixing exogenous and endogenous sequential decompositions. Non-uniform weights are particularly appropriate when there are substantive reasons to favor certain sequential decompositions. If a researcher believes one sequence of policy interventions is more plausible or relevant than another, they could use weights that reflect this prior belief. This flexibility highlights the core issue: without a substantive reason for choosing among them, the number of decompositions that are plausibly causal is infinite, underscoring the fundamental ambiguity of multi-cause attribution.

The conceptual distinction between a decomposition and a causal decomposition arises only in settings with multiple causes. With only one binary cause, there is only one decomposition function and it is inherently causal. Here, the type reduces to $\tilde{\theta} = (\tilde{a}, \tilde{y}(\cdot))$, for binary treatment $\tilde{a} \in \{0, 1\}$ and potential outcome function $\tilde{y}(\cdot) : \{0, 1\} \to \mathbb{R}$. The combined effect is the difference in potential outcomes, $\tau(\tilde{\theta}) = \tilde{y}(\tilde{a}) - \tilde{y}(0)$. Any decomposition must assign this entire quantity to the sole cause, meaning there is only one possible rule: $\psi(\tilde{\theta}) = \tau(\tilde{\theta})$. While practical challenges of identification and estimation remain, the conceptual task is unambiguous.

This clarity is lost when a second cause is introduced. The single, unambiguous answer with one cause gives way to a multiplicity of principled, yet conflicting, decompositions. The following subsections characterize this space of causal decompositions, establishing that the resulting ambiguity is fundamental.

2.3. A Constructive Definition of Causal Decompositions

This section develops the constructive argument that underpins the proposed definition of a causal decomposition. First, I deconstruct the causes' combined effect into its constituent causal pathways via a simple algebraic identity. This representation reveals a natural partition of the combined effect into components that are unambiguously attributable to a single cause and those that are jointly-produced by both causes, and thus ambiguous. This structure leads directly to our definition of a causal decomposition as a rule that assigns each cause its individually produced effects plus a share of its jointly-produced effects. The section concludes by illustrating how existing methods fit into this framework: while some correspond to specific causal decompositions, others do not.

To understand how to attribute the combined effect, $\tau(\theta)$, it helps to first understand

its structure. The combined effect is not a single, indivisible quantity. Rather, it is composed of a few primitive causal relationships: the standalone effect of each cause on the outcome ($y(1,0) - y(0,0)$ and $y(0,1) - y(0,0)$), the non-additive interaction between them ($y(1,1) - y(0,1) - (y(1,0) - y(0,0))$), and the effect of the first cause on the second ($a_2(1) - a_2(0)$). The following lemma shows that the combined effect can be algebraically rewritten as the sum of five distinct components, each representing a product of these primitive effects.

LEMMA 1. *For type, $\theta \in \Theta$, the combined effect, $\tau(\theta)$, can always be written as the sum of the following five component effects:*

$$\tau(\theta) = D_1(\theta) + D_2(\theta) + I(\theta) + X_D(\theta) + X_I(\theta)$$

where each component is defined as:

$$D_1(\theta) = \left[y(1,0) - y(0,0)\right] a_1 \qquad \textbf{\textit{1st Direct Effect}}$$

$$D_2(\theta) = \left[y(0,1) - y(0,0)\right] a_2(0) \qquad \textbf{\textit{2nd Direct Effect}}$$

$$I(\theta) = \left[y(0,1) - y(0,0)\right] (a_2(1) - a_2(0))\, a_1 \qquad \textbf{\textit{Indirect Effect}}$$

$$X_D(\theta) = \left[y(1,1) - y(0,1) - (y(1,0) - y(0,0))\right] a_2(0) a_1 \qquad \textbf{\textit{Direct Interaction}}$$

$$X_I(\theta) = \left[y(1,1) - y(0,1) - (y(1,0) - y(0,0))\right] (a_2(1) - a_2(0))\, a_1 \qquad \textbf{\textit{Indirect Interaction}}$$

PROOF. All proofs for this section are provided in Appendix A □

The first step in building the definition is to sort these five effects into two groups based on their source: individually-produced effects, which isolate the impact of a single cause, and jointly-produced effects, which depend on both. The two direct effects, $D_1(\theta)$ and $D_2(\theta)$, are individually-produced, as they isolate the effect of each cause in a world without the other. In contrast, the remaining three effects, $I(\theta)$, $X_D(\theta)$ and $X_I(\theta)$, are jointly-produced as they depend on both causes. For instance, the indirect effect, $I(\theta)$, requires both an effect of cause 1 on cause 2 and an effect of cause 2 on the outcome. Similarly, the direct interaction effect, $X_D(\theta)$, represents the incremental, synergistic effect of having both causes present, above and beyond their individual contributions. The indirect interaction effect, $X_I(\theta)$, in turn, requires both the same synergistic effect and an effect of cause 1 on cause 2. Each jointly-produced effect is fundamentally the result of both causes.

Individually-produced effects present no attributional ambiguity: each cause deserves full credit for the effects it produces alone. The direct effect of the first cause,

$D_1(\theta)$, isolates the effect of the first cause in a world where the second cause is absent. This is effectively the case of a single cause, where there is no ambiguity: this effect is due entirely to the first cause. The same logic applies to the second direct effect, $D_2(\theta)$, which isolates the effect of the second cause in a world without the first: this effect is due entirely to the second cause.

In contrast, jointly-produced effects are at the heart of the ambiguity because each cause has a defensible, competing claim to credit. For the indirect effect, $I(\theta)$, one could credit the initial (first) cause for setting the chain of events in motion or the proximate (second) cause for directly affecting the outcome. For the direct interaction effect, $X_D(\theta)$, one could argue the first cause deserves the credit because removing it would eliminate the interaction effect, yet an identical argument assigns credit entirely to the second cause. The indirect interaction effect, $X_I(\theta)$, combines these issues: one must contend with how to share credit for both the interaction component and the indirect pathway through which it operates.

This logic provides a simple recipe for constructing a causal decomposition: assign the individually-produced effects entirely to their sole cause, while allowing for any split of the jointly-produced ones. The following definition formalizes this.

DEFINITION 4. *A **causal decomposition** (or attribution rule), ψ, is a decomposition, $\psi \in \Psi$, that assigns to each cause its full direct effect, plus a share of each of the three jointly-produced effects, governed by the sharing parameters $(\lambda_1, \lambda_2, \lambda_3) \in [0,1]^3$:*

$$\begin{aligned}\psi_1(\theta) &= D_1(\theta) \quad + \lambda_1 I(\theta) \quad + \lambda_2 X_D(\theta) \quad + \lambda_3 X_I(\theta)\\ \psi_2(\theta) &= D_2(\theta) + (1-\lambda_1) I(\theta) + (1-\lambda_2) X_D(\theta) + (1-\lambda_3) X_I(\theta)\end{aligned}$$

*Let Ψ^c denote the set of causal decompositions. The output of a causal decomposition is called an **attribution value**.*

This definition of a causal decomposition, captures a wide set of decompositions grounded in the causal pathways that come together to create the combined effect. Moreover, it emphasizes that the fundamental ambiguity in attribution arises from an ambiguity in how to share credit for jointly-produced effects. In the next subsection, I complement this constructive argument with a formal one. There, I show that this same set of decompositions is uniquely characterized by four fundamental properties, providing a dual justification for this definition.

With this characterization in hand, I now reexamine how existing attribution methods fit within the framework. As the following examples from our earlier discussion

illustrate, some common methods are causal decompositions, each representing a specific choice of the sharing parameters, λ, while others are not.

EXAMPLE 1 (Last Touch Attribution, continuing from p. 9). *Last Touch Attribution is not a causal decomposition: no choice of sharing parameters $\lambda(\theta) \in [0,1]^3$ can reproduce it. Crucially, even if the only non-zero effect is the first cause's direct effect, $D_1(\theta)$, the second cause could receive all of the credit, violating the core principle that individually-produced effects should be assigned entirely to the cause that generated them.*

EXAMPLE 2 (Sequential Decompositions, continuing from p. 10). *Each of the four sequential decompositions we introduced are causal decompositions. Depending on the sequence being considered, the exogenous sequential decompositions correspond to sharing parameters $\lambda^{exogSD,1} = (0,1,1)$ or $\lambda^{exogSD,2} = (0,0,0)$ as they always give the indirect effect to the second cause. Similarly, the endogenous sequential decompositions correspond to sharing parameters $\lambda^{endogSD,1} = (1,1,1)$ or $\lambda^{endogSD,2} = (0,0,0)$ as all of the jointly-produced effects can go entirely to either cause. The choice of exogenous versus endogenous sequential decompositions corresponds to choosing which cause gets credit for the indirect effect, while the sequence one chooses determines which cause gets credit for the interaction effects.*

EXAMPLE 3 (Causal Shapley Values, continuing from p. 11). *Both Shapley values are causal decompositions. exogenous Shapley values correspond to sharing parameter, $\lambda^{exogSV} = (0, 1/2, 1/2)$, as the indirect effect goes fully to the second cause, while the interaction effects are split equally. In contrast, the endogenous Shapley value corresponds to the sharing parameters, $\lambda^{endogSV} = (1/2, 1/2, 1/2)$, as each jointly-produced effect is split equally between both causes.*

Ultimately, this characterization provides a common language for discussing and comparing different approaches to attribution. By parameterizing attribution in terms of the different ways one could assign credit for jointly-produced effects, this characterization promotes transparency: a researcher who prefers a specific decomposition is prompted to justify their choice of sharing parameters. This perspective also suggests a path toward drawing robust conclusions in the face of ambiguity: report the full range of credit assigned according to different causal decompositions. These bounds transparently communicate the fundamental ambiguity inherent in the problem. I formalize the concept of attribution bounds in Section 3.

2.4. A Formal Foundation for Causal Decompositions

While the constructive definition of a causal decomposition (Definition 4) has a strong, intuitive appeal, this section provides a complementary, formal justification. The ar-

gument proceeds in two steps. First, I propose three properties — Linearity, No Credit for No Effect, and No Blame for No Harm — and show that they characterize the set of causal decompositions for non-defiers (where the potential treatments are weakly increasing). Second, I introduce a final property, Symmetry of Causal Changes, that extends the characterization to the more complex case of defiers.

The first property, Linearity, is a standard regularity condition ensuring that a decomposition behaves as expected when outcomes are scaled or combined. It has two intuitive components. First, if we have two outcomes for a given unit (e.g., expenditure on fruit and vegetables), assigning credit for their sum should be equivalent to the sum of their individual credit assignments. Second, if an outcome is scaled by a constant (e.g., changing units from dollars to cents), the credit assigned should scale by the same constant.

PROPERTY 1 (Linearity). *A decomposition, $\psi \in \Psi$, satisfies Linearity if for every scalar $c \in \mathbb{R}$, type, $\theta = (a_1, a_2(\cdot), y(\cdot)) \in \Theta$, and potential outcome function, $\tilde{y}(\cdot) \in \mathcal{Y}$, we have:*

(i) Additivity: $\psi(a_1, a_2(\cdot), y(\cdot) + \tilde{y}(\cdot)) = \psi(a_1, a_2(\cdot), y(\cdot)) + \psi(a_1, a_2(\cdot), \tilde{y}(\cdot))$

(ii) Homogeneity of Degree 1: $\psi(a_1, a_2(\cdot), c \times y(\cdot)) = c \times \psi(a_1, a_2(\cdot), y(\cdot))$

The next two properties restrict the behavior of causal decompositions based on the presence and sign of causal effects. To formalize these properties, I first define the following marginal effects.

DEFINITION 5. *For $t = 1, 2$, the t^{th} causes' **marginal effects** are differences in potential outcomes resulting from a single, incremental change in the t^{th} treatment, holding other treatments fixed. For $\theta = (a_1, a_2(\cdot), y(\cdot)) \in \Theta$, these are:*

- *First cause's marginal effect:* $m_1(b_2; \theta) = y(1, b_2) - y(0, b_2)$ *for* $b_2 \in \{0, 1\}$
- *Second cause's marginal effect:* $m_2(b_1; \theta) = y(b_1, 1) - y(b_1, 0)$ *for* $b_1 \in \{0, 1\}$
- *Marginal indirect effect:* $m_{1\to2}(b_1; \theta) = y(b_1, a_2(1)) - y(b_1, a_2(0))$ *for* $b_1 \in \{0, 1\}$

The second property, No Credit for No Effect, states that a cause should receive zero credit if it does nothing, either because its treatment cannot be realized or because it has no marginal effect.

PROPERTY 2 (No Credit for No Effect). *A decomposition, $\psi \in \Psi$, satisfies No Credit for No Effect if for every $\theta = (a_1, a_2(\cdot), y(\cdot)) \in \Theta$:*

(i) If Cause 1 is inactive ($a_1 = 0$) or has no marginal effects, $m_1(\cdot\,;\theta) = m_{1\to 2}(\cdot\,;\theta) = 0$, it receives no credit: $\psi_1(\theta) = 0$.

(ii) If Cause 2 cannot occur ($a_2(0) = a_2(1) = 0$) or has no marginal effects, $m_2(\cdot\,;\theta) = 0$, it receives no credit: $\psi_2(\theta) = 0$.

The third property, No Blame for No Harm, formalizes the intuition that a cause should receive non-negative credit if it is always helpful. A cause is always helpful if its marginal effects and indirect marginal effects are all weakly positive and increasing.

PROPERTY 3 (No Blame for No Harm). *A decomposition, $\psi \in \Psi$, satisfies No Blame for No Harm, if for every $\theta \in \Theta$:*

(1) $m_1(1;\theta) \geq m_1(0;\theta) \geq 0$ *and* $m_{1\to 2}(1;\theta) \geq m_{1\to 2}(0;\theta) \geq 0$, *implies* $\psi_1(\theta) \geq 0$

(2) $m_2(1;\theta) \geq m_2(0;\theta) \geq 0$ *and* $m_{1\to 2}(1;\theta) \geq m_{1\to 2}(0;\theta) \geq 0$, *implies* $\psi_2(\theta) \geq 0$

For the class of non-defiers, types where the first treatment does not inhibit the second treatment ($a_2(1) \geq a_2(0)$), these three properties precisely characterize the set of causal decompositions.

PROPOSITION 1. *For a decomposition, $\psi \in \Psi$, the following two conditions are equivalent on the set of non-defiers, $\{(a_1, a_2(\cdot), y(\cdot) : a_2(1) \geq a_2(0)\}$:*

(i) ψ *satisfies Linearity, No Credit for No Effect, and No Blame for No Harm (Properties 1, 2, and 3)*

(ii) ψ *is a causal decomposition,* $\psi \in \Psi^c$

This result provides a strong foundation, but notably excludes defiers, where the first treatment prevents the second treatment ($a_2(1) < a_2(0)$). Relying on only these three properties allows for an undesirable inconsistency in the underlying logic of attribution: defiers and non-defiers could have different sharing parameters. For example, a decomposition might credit the first cause for an indirect effect among compliers, while crediting the second cause among defiers. This is internally inconsistent: if a decomposition's logic is to blame the first cause for inhibiting an effect among defiers, that same logic should lead it to reward the first cause for instigating it among compliers.

To enforce this logical consistency, I introduce a final property: Symmetry of Causal Changes. This property says that enabling a complier causal pathway, should be rewarded the same as disabling an inhibitory causal pathway.

PROPERTY 4 (Symmetry of Causal Changes). *A decomposition, $\psi \in \Psi$, satisfies Symmetry of Causal Changes if for any $y(\cdot) \in \mathcal{Y}$ and $a_1 \in \mathcal{A}_1$, the difference in credit between a complier ($a_2^C(x_1) = x_1$) and a never-taker ($a_2^{NT}(x_1) = 0$), is equivalent to the difference in credit between an always taker ($a_2^{AT}(x_1) = 1$) and a defier ($a_2^D(x_1) = 1 - x_1$):*

$$\psi\left(a_1, a_2^C(\cdot), y(\cdot)\right) - \psi\left(a_1, a_2^{NT}(\cdot), y(\cdot)\right) = \psi\left(a_1, a_2^{AT}(\cdot), y(\cdot)\right) - \psi\left(a_1, a_2^D(\cdot), y(\cdot)\right)$$

This property bridges the non-defier and defier worlds, so that these four properties taken together precisely characterize the set of causal decompositions for all types.

THEOREM 1. *A decomposition, $\psi \in \Psi$, satisfies Linearity, No Credit for No Effect, No Blame for No Harm, and Symmetry of Causal Changes if and only if it is a causal decomposition, $\psi \in \Psi^c$.*

This theorem provides the formal counterpart to the constructive argument. It establishes that the structure of a causal decomposition in Definition 4 is not merely an intuitive choice, but also a necessary consequence of four properties a causal decomposition ought to satisfy.

The main implication of Definition 4 and Theorem 1 is that the ambiguity in attribution is fundamental and is parameterized by the choice of sharing parameters, $(\lambda_1, \lambda_2, \lambda_3)$. This insight shifts the research objective: instead of seeking a single correct answer, the goal becomes transparently reporting the range of plausible answers. To this end, I advocate for reporting attribution bounds: the minimum and maximum credit assigned to each cause across all causal decompositions. This practice makes the inherent ambiguity explicit. Moreover, the framework provides a common language for researchers who wish to narrow this range by allowing them to translate their substantive justifications directly into transparent restrictions on the sharing parameters. The following section formalizes these bounds and discusses how they can be identified and estimated in practice.

3. Identification and Estimation of Attribution Bounds

This section introduces a statistical framework for defining, identifying, and estimating attribution bounds. I define the estimand, attribution bounds, as bounds on the set of average attribution values. I then present two identification strategies: one design-based and one observational. The first result is negative: attribution bounds cannot

be identified by design alone. As a solution, I propose a simple condition on the signs of the average joint effects which, when met, enables identification by design. I show that this design-based approach hinges on the ability to randomize eligibility to receive treatment, rather than randomizing treatment itself. In contrast, identification is more straightforward in observational settings, at the expense of relying on a stronger assumption. I demonstrate that under a standard sequential unconfoundedness assumption (Imai et al. 2010b), the attribution bounds are point-identified. For both settings, I provide estimators and methods for constructing confidence intervals.

3.1. Statistical Setup and Experimental Design

This section introduces a statistical framework that incorporates Holland's fundamental problem of causal inference — missing potential outcomes — and statistical uncertainty (Holland 1986). In the process, I formalize the experimental instruments and the causal model that link observed data to latent potential outcomes and potential treatments.

Suppose a random sample of N individuals ($i = 1, \ldots, N$) is drawn from an infinitely large population. For each individual i, the realized outcome, Y_i, and sequence of treatments, $A_i = (A_{i1}, A_{i2}) \in \mathbb{A}$, are observed. As before, each individual has a latent potential outcome function, $Y_i(\cdot) : \mathbb{A} \to \mathbb{R}$, but now the data-generating process for treatments is explicitly modeled to incorporate their dependence on experimental interventions.

To formalize our experimental design, I introduce two experimental instruments. First, individuals are assigned indicators, $F_i = (F_{i1}, F_{i2}) \in \{0, 1\}^2 \equiv \mathbb{F}$, determining if they receive treatment endogenously ($F_{it} = 1$) or exogenously ($F_{it} = 0$). Second, they are assigned an exogenous treatment sequence, $A_i^{\mathrm{s}} = (A_{i1}^{\mathrm{s}}, A_{i2}^{\mathrm{s}}) \in \mathbb{A}$.[5] These experimental instruments jointly determine individual i's realized treatment, A_{it}, in each period t. When treatment is endogenous ($F_{it} = 1$), the individual's realized treatment is determined by their own potential treatment function, representing their natural behavior. When treatment is exogenous ($F_{it} = 0$), this behavior is overridden, and their realized treatment is set to A_{it}^{s}. For example, $F_i = (0, 0)$ with $A_i^{s} = (0, 0)$ indicates that both treatments are to be exogenously set to 0, a pure control, while $F_i = (1, 1)$ corresponds to observational data.

The role of these experimental instruments necessitates new notation for potential treatments. Let $A_{i1}(1)$ denote the treatment individual i would endogenously receive in period 1 (i.e., if $F_{i1} = 1$). This is their baseline behavior in the first period. For the second

[5]The superscript 's' can be read as 'set', denoting the treatment 'set' by the experimenter when $F_{it} = 0$.

cause, the potential treatment can depend on the first treatment. Let $A_{i2}(a_1, 1)$ represent the treatment individual i would receive in the second period if the first treatment had been $a_1 \in \mathbb{A}_1$ and the second-period treatment were endogenous (i.e., if $F_{i2} = 1$). This notation highlights an asymmetry: potential treatments are defined only when treatment is endogenous ($F_{it} = 1$). This is because potential variables represent an immutable, intrinsic characteristic of an individual. In contrast, an exogenous assignment, like A^s_{it}, is a feature of the experimental protocol, not a latent attribute of the individual.

The following assumption formalizes the relationships between potential variables, experimental instruments, and observed data.

ASSUMPTION 1 (Causal Model). *For each individual i, the observed outcome Y_i and observed treatments A_i are generated as follows:*

$$\begin{aligned} Y_i &= Y_i(A_i) \\ A_{i1} &= A_{i1}(1)F_{i1} + A^s_{i1}(1 - F_{i1}) \\ A_{i2} &= A_{i2}(A_{i1}, 1)F_{i2} + A^s_{i2}(1 - F_{i2}) \end{aligned}$$

where the potential outcomes function $Y_i(\cdot)$ and potential treatment function $A_{it}(\cdot)$ are understood to be intrinsic properties of the individual that do not depend on the assignments of other individuals.

Assumption 1 bundles three foundational concepts. First, it states the *consistency* of outcomes and treatments: the observed variables are equal to their corresponding potential counterparts evaluated at the realized treatment and instrument values. Second, it implies *no cross-unit spillovers* (no interference) by defining potential outcomes and treatments for individual i without reference to the treatments or assignments of others. Third, it encodes an *exclusion restriction*: the experimental instruments (F_{it}, A^s_{it}) affect outcomes and future treatments only through their effect on the realized treatment in that period, A_{it}.

3.2. The Estimand: Attribution Bounds

Here, I define the primary estimand: attribution bounds. I then express the attribution bounds as the minimum/maximum of a linear combination of six counterfactual quantities. This representation reveals an important identification challenge: two of these quantities are cross-world counterfactuals, which are unidentifiable from experimental data alone.

This paper introduces attribution bounds as its primary estimand. These bounds are designed to communicate the full range of expected attribution values, $\mathbb{E}[\psi_t(A_1, A_2(\cdot), Y(\cdot))]$, that can arise from different causal decompositions, $\psi \in \Psi^c$.

DEFINITION 6. *The **attribution bounds** for the t^{th} cause, $[\underline{V}_t, \overline{V}_t]$, are the minimum and maximum expected attribution values over the set of all valid causal decompositions, $\psi \in \Psi^c$:*

$$\underline{V}_t \equiv \min_{\psi \in \Psi^c} \mathbb{E}\left[\psi_t\left(A_1, A_2(\cdot), Y(\cdot)\right)\right] \quad \textit{and} \quad \overline{V}_t \equiv \max_{\psi \in \Psi^c} \mathbb{E}\left[\psi_t\left(A_1, A_2(\cdot), Y(\cdot)\right)\right]$$

The attribution bounds, $[\underline{V}_t, \overline{V}_t]$, represent a set of estimands, not an identified set for a single estimand. This is a critical distinction. An identified set reflects the range of answers the data cannot distinguish between for a single question. In contrast, attribution bounds reflect the ambiguity in the question: each causal decomposition represents a different causal question. Indeed, the data may be sufficient to point-identify the answer to every such question, even though they cannot dictate which question is the right one to ask.

The characterization of causal decompositions (Definition 4) makes it possible to express the attribution bounds as the solution to a simple linear optimization problem.

LEMMA 2. *For $t \in \{1, 2\}$, the attribution bounds, $[\underline{V}_t, \overline{V}_t]$, are the minimum and maximum values of the following function evaluated over the set $\lambda = (\lambda_1, \lambda_2, \lambda_3) \in [0, 1]^3$:*

$$\begin{aligned}
V_t(\lambda) = {} & \mathbb{E}[Y(0,0)] \times (\lambda_2 - 1) && (i)\\
& + \mathbb{E}[Y(A_1(1), 0)] \times (\mathbf{1}\{t = 1\} - \lambda_2) && (ii)\\
& + \mathbb{E}[Y(0, A_2(0,1))] \times (\mathbf{1}\{t = 2\} - \lambda_1 - \lambda_2 + \lambda_3) && (iii)\\
& + \mathbb{E}[Y(0, A_2(A_1(1), 1))] \times (\lambda_1 - \lambda_3) && (iv)\\
& + \mathbb{E}[Y(A_1(1), A_2(0,1))] \times (\lambda_2 - \lambda_3) && (v)\\
& + \mathbb{E}[Y(A_1(1), A_2(A_1(1), 1))] \times \lambda_3 && (vi)
\end{aligned}$$

PROOF. All proofs for this section are provided in Appendix B □

This lemma highlights that identification hinges on six counterfactual quantities. Moreover, it reveals why the bounds are not identified by design alone: terms (iv) and (v) are cross-world counterfactuals. Term (iv), $\mathbb{E}[Y(0, A_2(A_1(1), 1))]$, for instance, describes an outcome where the first treatment is set exogenously to 0, but where the

second treatment occurs as if the first treatment had been endogenous ($A_1(1)$). The very act of enforcing the first condition via an experimental manipulation renders the second condition unobservable, ensuring these terms cannot be identified by design. To juxtapose the cross-world counterfactuals, I refer to the other four terms (i, ii, iii, and vi) as single-world counterfactuals because they correspond to outcomes from a single counterfactual. For instance, term (iii), $\mathbb{E}[Y(0, A_2(0, 1))]$, describes an outcome and potential treatment where the first treatment is set exogenously to 0 and the treatment for the second cause is determined endogenously ($F_{i2} = 1$).

The following subsections propose two different strategies for overcoming this difficulty. First, to achieve design-based identification, Section 3.3 proposes a simple assumption on the sign of the jointly-produced effects that, if satisfied, ensures the attribution bounds do not depend on the cross-world counterfactuals. Second, in observational settings (Section 3.4), I consider a stronger assumption, sequential unconfoundedness, which identifies all of the relevant counterfactuals, including the cross-world ones.

3.3. Design-Based Identification, Estimation and Inference

This section introduces a general class of experimental designs and shows that no design within this class can identify the attribution bounds without further assumptions. However, I show that if the average jointly-produced effects have the same sign, the attribution bounds can be identified by design. I characterize the minimal class of experiments, partial eligibility designs, that are sufficient for identification under the proposed condition. Moreover, I show that a subset of partial eligibility designs, which I call eligibility designs, are weakly more informative when the true bounds are non-trivial (that is, when they do not collapse). Finally, I provide estimators and confidence intervals for attribution bounds estimated by a partial eligibility experiment.

3.3.1. Identification

To start, I introduce a broad space of experimental designs where the experimental instruments are randomly assigned and individuals have at least some probability of being intervened on.

DESIGN 1. *An **experimental design**, $P \in \mathcal{E} \subset \Delta(\mathbb{F} \times \mathbb{A})$, is a distribution over the experimental instruments, $(F_i, A_i^s) \sim P$, that (i) exogenously assigns treatment with a positive*

probability and (ii) is independent of potential variables:

$$\mathcal{E} = \Big\{ P \in \Delta(\mathbb{F} \times \mathbb{A}) \;:\; P(F = (1,1)) < 1\,, \\ \big(F, A^s\big) \perp \big(\{Y(a)\}_{a\in\mathbb{A}}\,,\, A_1(1)\,,\, \{A_2(a_1,1))\}_{a_1\in\mathbb{A}_1}\big) \Big\}$$

In this general class of designs, an experimenter can randomize both which treatments are free to vary endogenously, F_{it}, and what treatments they receive when they are not, A_i^s. Despite this flexibility, no design can point-identify the attribution bounds on its own because they depend on the cross-world counterfactuals (Lemma 2).

PROPOSITION 2. *Suppose Assumption 1 holds. There is no experimental design $P \in \mathcal{E}$ that point-identifies the attribution bounds.*

To resolve the identification challenge posed by the cross-world counterfactuals, I introduce a new assumption on the signs of the average jointly-produced effects. When met, this assumption ensures that the attribution bounds do not depend on the unidentifiable cross-world counterfactuals.

ASSUMPTION 2 (Sign-Aligned Joint Effects). *The average indirect effect, $\mathbb{E}[I(A_1, A_2(\cdot), Y(\cdot))]$, average direct interaction effect, $\mathbb{E}[X_D(A_1, A_2(\cdot), Y(\cdot))]$, and average indirect interaction effect, $\mathbb{E}[X_I(A_1, A_2(\cdot), Y(\cdot))]$ are all weakly of the same sign (i.e., they are all ≥ 0 or ≤ 0).*

Assumption 2 posits that the joint effects of the causes all go in the same direction. This condition holds, for instance, if the first cause always makes the second one more likely, the second cause on its own is always helpful, and both causes always have a positive interaction effect. Crucially, however, this condition applies to average effects, allowing for individual-level heterogeneity where some units may not conform to the overall trend. A special case where the assumption mechanically holds is when the first cause has no effect on the second. Conversely, the assumption would be violated if there was a positive indirect effect, but the causes are substitutes (i.e., have a negative interaction effect).

This assumption does not help us identify the cross-world counterfactuals. Rather, it ensures they drop out of the optimization problem in Lemma 2, making them irrelevant for the attribution bounds. Interestingly, this highlights that it can be easier to identify the attribution bounds than to identify the expected attribution values for a specific causal decomposition, which may still depend on these unidentifiable terms.

Even with Assumption 2, however, not all experimental designs generate the necessary variation to identify the attribution bounds. For instance, the class of experimental

designs is broad enough to include trivial designs that deterministically put everyone in a pure control group. This motivates defining a minimal class of experimental designs that achieve point-identification.

DESIGN 2. *A **partial eligibility design**, $P \in \mathcal{E}^{PE}$, is an experimental design that (a) has full support over which treatments can be endogenous responses and (b) has a positive probability of assigning individuals to receive no treatment in each period:*

$$\mathcal{E}^{PE} = \Big\{ P \in \mathcal{E} \ : P(F = f) > 0 \ \forall f \in \mathbb{F} \ , \\ P\left(A^S = (0,0) | F = (0,0)\right) > 0 \ , \\ P(A^S_t = 0 | F_t = 0, F_{-t} = 1) > 0 \ \forall t \in \{1,2\} \Big\}$$

The conditions of a partial eligibility design are precisely what is needed to identify the four single-world counterfactuals that determine the bounds under Assumption 2. Condition (a) ensures that with positive probability one observes outcomes under all four combinations of endogenous/exogenous treatments, while condition (b) ensures that with positive probability one can isolate the effects of endogenous behavior against a common baseline (no treatment). These support restrictions are both necessary and sufficient.

PROPOSITION 3. *Under Assumptions 1 and 2, an experimental design, $P \in \mathcal{E}$, point identifies the attribution bounds if and only if it is a partial eligibility design, $P \in \mathcal{E}^{PE}$.*

While all partial eligibility designs are sufficient for identification, some are better than others. Here, I characterize a subclass that is optimal whenever the bounds are non-trivial (i.e., $\underline{V}_t < \overline{V}_t$).

DESIGN 3. *An **eligibility design**, $P \in \mathcal{E}^{E}$, is a partial eligibility design that assigns individuals to the control group, $A_{it} = 0$, whenever treatment is exogenously set, $F_{it} = 0$. Formally:*

$$\mathcal{E}^{E} = \Big\{ P \in \mathcal{E}^{PE} \ : \ P\left(A^S_t = 0 | F_t = 0\right) = 1 \, \forall t \Big\}$$

An eligibility design is conceptually simple: it randomizes eligibility to receive treatment. In each period, an individual is either eligible to receive treatment and does so as a function of their potential treatments ($F_{it} = 1$), or they are ineligible to receive treatment ($F_{it} = 1$, $A^S_{it} = 0$). This experiment isolates the variation needed for identifying the attribution bounds under Assumption 2 by only allocating units to one of the four

relevant, single-world counterfactual means. The following proposition clarifies the sense in which eligibility designs are preferable to partial-eligibility designs.

PROPOSITION 4. *Let $P \in \mathcal{E}^{PE}$ be any partial-eligibility design. There exists an eligibility design $P' \in \mathcal{E}^{E}$ that is weakly more efficient for estimating the four single-world counterfactuals.*[6] *Moreover, if the true bounds are non-trivial ($\underline{V}_t < \overline{V}_t$), this efficiency gain extends to the estimation of the attribution bounds.*

Intuitively, any other partial eligibility design expends subjects on experimental arms that are not informative of the single-world counterfactuals. By allocating all exogenous treatments to control, eligibility designs can achieve lower asymptotic variance for these means. The gain in asymptotic variance for the attribution bounds follows directly from this, with one nuance. When the true bounds are distinct ($\underline{V}_t < \overline{V}_t$), the bounds can be estimated by a smooth function of the efficiently estimated means, thus inheriting their efficiency. This argument fails only when the bounds collapse ($\underline{V}_t = \overline{V}_t$), as they are no longer a locally-differentiable function of the means, and standard asymptotic optimality arguments break down. Despite this nuance, eligibility designs are the clear practical choice: they are more efficient for estimating the single-world counterfactual means, which are often of independent interest, and for estimating the bounds across nearly the entire parameter space.

3.3.2. Estimation and Inference

Given data from a partial eligibility design, I show how one can estimate attribution bounds and provide methods for constructing confidence intervals. The estimation strategy is straightforward: first, estimate the four key counterfactual means using an inverse probability weighted (Horvitz-Thompson) estimator to account for the experimental design. Second, use these estimates to form a plug-in estimator for the bounds. I then characterize the non-normal asymptotic distribution of this estimator and provide a simulation-based algorithm for constructing asymptotically valid confidence intervals.

Under Assumption 2, the optimization problem in Lemma 2 simplifies, and the bounds become the minimum and maximum of two specific linear combinations of

[6] Formally, assuming finite variance for all potential outcomes, the asymptotic variance of any regular and asymptotically linear (RAL) estimator for the means is weakly lower under P' than under P. The class of RAL estimators includes standard estimators like the sample mean, OLS, and AIPW.

the four single-world counterfactuals means. These two linear combinations are characterized by, $\tilde{V}_t(\lambda)$ for $\lambda \in \{0, 1\}$:

$$\begin{aligned}\tilde{V}_t(\lambda) = \mathbb{E}[Y(0,0)] \times (\lambda - 1) \;+\; \mathbb{E}[Y(A_1(1),0)] \times (\mathbf{1}\{t=1\} - \lambda) \\ +\; \mathbb{E}[Y(0, A_2(0,1))] \times (\mathbf{1}\{t=2\} - \lambda) \;+\; \mathbb{E}[Y(A_1(1), A_2(A_{i1}(1),1))] \times \lambda\end{aligned}$$

The proposed estimators target these two quantities, $\tilde{V}_t(0)$ and $\tilde{V}_t(1)$, and reports the lesser (greater) of the two as estimates of the lower (upper) attribution bounds. Though the estimator will work with any partial eligibility design, the estimator's structure reveals that it only uses data from the four experimental arms corresponding to the single-world counterfactuals. This reinforces the efficiency of the eligibility design, which concentrates the entire sample on these informative arms.

ESTIMATOR 1 (Plug-In Estimator). *First, estimate the four single-world counterfactual means, μ_f for $f \in \mathbb{F}$. For each mean, define an individual inverse probability weighted (IPW) outcome, $Z_{i,f}$, as:*

$$Z_{i,f} = \frac{\mathbf{1}\{F_i = f\,,\; A_i^s \cdot (1 - F_i) = 0\}}{P(F_i = f\,,\; A_i^s \cdot (1 - F_i) = 0)} Y_i$$

The indicator in the numerator isolates individuals assigned to the experimental arm corresponding to endogenous treatment assignments, $f \in \mathbb{F}$, and who are exogenously assigned to control whenever treatment is exogenously assigned. The denominator reweighs the outcome by the known assignment probability. Then the Horvitz-Thompson estimator for each mean is the sample average of these weighted outcomes: $\hat{\mu}_f = N^{-1} \sum_i Z_{i,f}$.

These estimated means and IPW outcomes are collected in vectors: $\hat{\mu} = (\hat{\mu}_{00}\,,\; \hat{\mu}_{10}\,,\; \hat{\mu}_{01}\,,\; \hat{\mu}_{11})^T$ and $Z_i = (Z_{i,00}\,,\; Z_{i,10}\,,\; Z_{i,01}\,,\; Z_{i,11})^T$. Then, the estimated means are used to form plug-in estimates of the two candidate values for the bounds, $(\tilde{V}_t(0), \tilde{V}_t(1)) = \Lambda_t \hat{\mu}$, where Λ_t is the transformation matrix:

$$\Lambda_t = \begin{pmatrix} -1 & \mathbf{1}\{t=1\} & \mathbf{1}\{t=2\} & 0 \\ 0 & \mathbf{1}\{t=1\} - 1 & \mathbf{1}\{t=2\} - 1 & 1 \end{pmatrix}$$

Finally, the estimated attribution bounds are the minimum and maximum of these two values ($\Lambda_t \hat{\mu}$):

$$\underline{\hat{V}}_t = \min\, \Lambda_t \hat{\mu} \quad \& \quad \hat{\overline{V}}_t = \max\, \Lambda_t \hat{\mu}$$

Next, it is shown that this estimator is consistent, but because of the estimator's min / max operations, it converges to a non-normal distribution.

THEOREM 2. *The estimators, $\underline{\hat{V}}_t$ and $\hat{\overline{V}}_t$, are consistent. Furthermore, under standard regularity conditions for the central limit theorem, $\sqrt{N}(\Lambda_t\hat{\mu} - \Lambda_t\mu)$ converges to a multivariate normal distribution, $\mathcal{N}\left(0, \Lambda_t\Sigma\Lambda_t^T\right)$, as N grows, where $\Sigma = Cov(Z_i)$ and $\mu = \mathbb{E}[Z_i]$. The bounds themselves, $\underline{\hat{V}}_t$ and $\hat{\overline{V}}_t$, converges either to the min/max of two normals, $\Lambda_t\hat{\mu}$, when the bounds collapse (i.e., $\underline{V}_t = \overline{V}_t$), or to a normal centered around the corresponding bound when the bounds do not collapse (i.e., $\underline{V}_t < \overline{V}_t$).*

Due to the non-linearity of the min / max operators, the estimators, $\underline{\hat{V}}_t$ and $\hat{\overline{V}}_t$, are biased. By Jensen's inequality, this bias is conservative, the estimated bounds are, on average, wider than the true bounds: $\mathbb{E}[\underline{\hat{V}}_t] \leq \underline{V}_t$ and $\overline{V}_t \leq \mathbb{E}[\hat{\overline{V}}_t]$.

To construct confidence intervals, I propose a subsampling-based approach (Politis and Romano 1994). While the bootstrap is a more common resampling method, it is known to fail when the estimator is not locally differentiable, as is the case for $\underline{V}_t$ and $\overline{V}_t$ when the true bounds collapse ($\underline{V}_t = \overline{V}_t$). In contrast, subsampling is theoretically guaranteed to provide asymptotically valid confidence intervals even in such non-standard settings. Algorithm 1 formalizes this robust inference procedure.

3.4. Identification by Sequential Unconfoundedness

When experimentation is infeasible, identification can be achieved by relying on a sequential unconfoundedness assumption. This approach, common in observational studies with time-varying treatments, requires that treatment assignment in each period is as good as random, conditional on the observed history. This section shows how this assumption identifies all six counterfactual means needed for the attribution bounds, and provides a corresponding plug-in estimator and a bootstrap-based method for constructing confidence intervals.

In a purely observational setting, all individuals follow their natural treatment path, which corresponds to the case where everyone endogenously receives treatment, i.e., $F_i = (1, 1)$ for all i. Given that there is no meaningful variation for the experimental instruments in an observational study, the potential treatment notation can be simplified, letting $A_{i1} = A_{i1}(1)$ and $A_{i2}(a_1) = A_{i2}(a_1, 1)$, such that the observed treatment path is $A_i = (A_{i1}, A_{i2}(A_{i1}))$.

[7] Note: The subsample size b must satisfy $b \to \infty$ and $b/N \to 0$. A common practical choice is $b = \lfloor N^c \rfloor$ for some $c \in (0.5, 1)$, such as $c = 4/5$.

Algorithm 1 Subsampling Confidence Intervals for Attribution Bounds

Input: Data $D = \{Y_i, F_i, A_i^s\}_{i=1}^N$; subsample size b;[7] number of subsamples S; confidence level α.

Output: $(1 - \alpha)$ confidence intervals for $\underline{V}_t$ and $\overline{V}_t$.

1: Compute the full-sample estimates $\hat{\underline{V}}_t$ and $\hat{\overline{V}}_t$ from D using Estimator 1.
2: **for** $s = 1$ **to** S **do**
3: Draw a subsample D_s^* of size b from D without replacement.
4: Compute the subsample estimates $\hat{\underline{V}}_{t,s}^*$ and $\hat{\overline{V}}_{t,s}^*$ from D_s^* using Estimator 1.
5: Define the s^{th} centered statistics:

$$q_{\underline{V},s} \leftarrow \sqrt{b}\left(\hat{\underline{V}}_{t,s}^* - \hat{\underline{V}}_t\right)$$
$$q_{\overline{V},s} \leftarrow \sqrt{b}\left(\hat{\overline{V}}_{t,s}^* - \hat{\overline{V}}_t\right)$$

6: **end for**
7: Let $c_{\underline{V}}(\beta)$ be the β-quantile of the empirical distribution of $\{q_{\underline{V},s}\}_{s=1}^S$.
8: $CI_{\underline{V}} \leftarrow \left[\hat{\underline{V}}_t - \frac{c_{\underline{V}}(1-\alpha/2)}{\sqrt{N}}, \quad \hat{\underline{V}}_t - \frac{c_{\underline{V}}(\alpha/2)}{\sqrt{N}}\right]$.
9: Let $c_{\overline{V}}(\beta)$ be the β-quantile of the empirical distribution of $\{q_{\overline{V},s}\}_{s=1}^S$.
10: $CI_{\overline{V}} \leftarrow \left[\hat{\overline{V}}_t - \frac{c_{\overline{V}}(1-\alpha/2)}{\sqrt{N}}, \quad \hat{\overline{V}}_t - \frac{c_{\overline{V}}(\alpha/2)}{\sqrt{N}}\right]$.
11: **return** $CI_{\underline{V}}, CI_{\overline{V}}$

To substitute for the lack of randomization, I rely on the following sequential unconfoundedness assumption, also known as sequential ignorability (Imai et al. 2010b; Hernán and Robins 2020). This assumption requires that treatment in each period is independent of all potential future outcomes and behaviors conditional on the observed history; the observed history includes past treatments and a vector of pre-treatment covariates, $X_i \in \mathbb{X} \subset R^d$.

ASSUMPTION 3 (Sequential Unconfoundedness). *For all $a_1, \tilde{a}_1 \in \mathbb{A}_1$, $a_2 \in \mathbb{A}_2$, and $x \in \mathbb{X}$, we have:*

1. $(Y(a_1, a_2), A_2(\tilde{a}_1)) \perp A_1 | X = x$
2. $Y(a_1, a_2) \perp A_2(\tilde{a}_1) | A_1 = \tilde{a}_1, X = x$

The first condition states that the initial treatment, A_1, is conditionally independent of all potential outcomes and future potential treatments. This is the standard unconfoundedness (or 'as-if random') assumption for a single treatment. The second

condition is stronger; it assumes that the second-period treatment is also conditionally independent of potential outcomes, even those under different treatment histories. This includes a cross-worlds independence assumption, which, while strong, is standard for achieving point identification in this context, and is precisely what we need to identify the cross-world counterfactual means.[8]

In addition to sequential unconfoundedness, there needs to be a positive probability that every individual receives every treatment combination. This is often called a positivity or overlap assumption:

ASSUMPTION 4 (Positivity). *For all* $(a_1, a_2) \in \mathbb{A}$ *and* $x \in \mathbb{X}$*, we have:*

$$\mathbb{P}(A_2 = a_2 | A_1 = a_1, X = x) > 0$$
$$\mathbb{P}(A_1 = a_1 | X = x) > 0$$

Under these assumptions, the attribution bounds are identified because all six, single-world and cross-world, counterfactual means outlined in Lemma 2 are themselves identified.

PROPOSITION 5. *Under Assumptions 1, 3 and 4, all six counterfactuals means required for the attribution bounds (Lemma 2) are identified and can be expressed in terms of observable distributions:*

$$\begin{aligned}
\mathbb{E}[Y(0,0)] &= \mathbb{E}\left[\mathbb{E}[Y|A=(0,0),X]\right] \\
\mathbb{E}[Y(A_1,0)] &= \sum_{a_1} \mathbb{E}\left[\mathbb{E}[Y|A=(a_1,0),X]\mathbb{P}(A_1=a_1|X)\right] \\
\mathbb{E}[Y(0,A_2(0))] &= \sum_{a_2} \mathbb{E}\left[\mathbb{E}[Y|A=(0,a_2),X]\mathbb{P}(A_2=a_2|A_1=0,X)\right] \\
\mathbb{E}[Y(0,A_2(A_1))] &= \sum_{a_2} \mathbb{E}\left[\mathbb{E}[Y|A=(0,a_2),X]\mathbb{P}(A_2=a_2|X)\right] \\
\mathbb{E}[Y(A_1,A_2(0))] &= \sum_{a_1,a_2} \mathbb{E}\left[\mathbb{E}[Y|A=(a_1,a_2),X]\mathbb{P}(A_1=a_1|X)\mathbb{P}(A_2=a_2|A_1=0,X)\right] \\
\mathbb{E}[Y(A_1,A_2)] &= \mathbb{E}[Y]
\end{aligned}$$

As a consequence, the attribution bounds, $\underline{V}_t$ *and* $\overline{V}_t$*, are also identified for each* t*.*

[8] As with the experimental setting, if the average joint effects have the same sign (Assumption 2), it is not necessary to identify the cross-world counterfactuals and thus the analysis does not need to rely on cross-worlds independence.

Algorithm 2 Cross-Fitted Plug-In Estimator for Attribution Bounds

Input: Data $D = \{Y_i, A_i, X_i\}_{i=1}^N$; number of folds K.

Output: Estimated attribution bounds $\hat{\underline{V}}_t, \hat{\overline{V}}_t$.

1: Randomly partition the set of indices $\{1, \ldots, N\}$ into K disjoint folds, $I_1, \ldots, I_K$.
2: **for** $k = 1$ **to** K **do**
3: Define the training data as $D^{(-k)} = D \setminus \{D_i\}_{i \in I_k}$.
4: Using $D^{(-k)}$, estimate nuisance functions: $\hat{\mu}^{(-k)}(a_1, a_2, x)$, $\hat{p}^{(-k)}(a_2|a_1, x)$, and $\hat{p}^{(-k)}(a_1|x)$.
5: **for** each observation $i \in I_k$ **do**
6: For all $(a_1, a_2) \in \mathbb{A}$, store $\hat{\mu}_i(a_1, a_2) \leftarrow \hat{\mu}^{(-k)}(a_1, a_2, X_i)$.
7: For all $(a_1, a_2) \in \mathbb{A}$, store $\hat{p}_i(a_2|a_1) \leftarrow \hat{p}^{(-k)}(a_2|a_1, X_i)$ and $\hat{p}_i(a_1) \leftarrow \hat{p}^{(-k)}(a_1|X_i)$.
8: **end for**
9: **end for**
10: Compute counterfactual means using predictions.

$$\hat{\mathbb{E}}[Y(0,0)] \leftarrow N^{-1} \sum_{i=1}^N \hat{\mu}_i(0,0).$$
$$\hat{\mathbb{E}}[Y(A_1,0)] \leftarrow N^{-1} \sum_{i=1}^N \sum_{a_1 \in \mathbb{A}_1} \hat{\mu}_i(a_1,0)\, \hat{p}_i(a_1).$$
$$\hat{\mathbb{E}}[Y(0,A_2(0))] \leftarrow N^{-1} \sum_{i=1}^N \sum_{a_2 \in \mathbb{A}_2} \hat{\mu}_i(0,a_2)\, \hat{p}_i(a_2|0).$$
$$\hat{\mathbb{E}}[Y(0,A_2(A_1))] \leftarrow N^{-1} \sum_{i=1}^N \sum_{a_2 \in \mathbb{A}_2} \hat{\mu}_i(0,a_2) \left(\sum_{a_1 \in \mathbb{A}_1} \hat{p}_i(a_2|a_1)\, \hat{p}_i(a_1) \right).$$
$$\hat{\mathbb{E}}[Y(A_1,A_2(0))] \leftarrow N^{-1} \sum_{i=1}^N \sum_{a_1,a_2} \hat{\mu}_i(a_1,a_2)\, \hat{p}_i(a_1)\, \hat{p}_i(a_2|0).$$
$$\hat{\mathbb{E}}[Y(A_1,A_2)] \leftarrow N^{-1} \sum_{i=1}^N Y_i.$$

11: Compute $\hat{V}_t(\lambda)$ by plugging $(\hat{\mathbb{E}}[Y(0,0)], \ldots, \hat{\mathbb{E}}[Y(A_1,A_2)])^T$ into the formula from Lemma 2.
12: $\hat{\underline{V}}_t \leftarrow \min_{\lambda \in \{0,1\}^3} \hat{V}_t(\lambda)$.
13: $\hat{\overline{V}}_t \leftarrow \max_{\lambda \in \{0,1\}^3} \hat{V}_t(\lambda)$.
14: **return** $\hat{\underline{V}}_t, \hat{\overline{V}}_t$

The identification results in Proposition 5 provides a path for estimation via a plug-in approach. This requires estimating two sets of nuisance functions: the conditional outcome function, $\mu(a_1, a_2, x) \equiv \mathbb{E}[Y|A = (a_1, a_2), X = x]$, and the sequential propensity scores, $p(a_2|a_1, x) \equiv \mathbb{E}[A_2|A_1 = a_1, X = x]$ and $p(a_1|x) \equiv \mathbb{E}[A_1|X = x]$. Here, we employ a K-fold cross-fitting procedure, which involves sample-splitting to estimate the nuisance functions and use them to generate predictions on separate, independent samples. The procedural nature of this approach is formalized in Algorithm 2.

The following theorem establishes the asymptotic properties of the cross-fitted estimator. While the intermediate counterfactual means are asymptotically normal,

the final bound estimators inherit a non-standard limiting distribution for analogous reasons to their design-based counterparts (Theorem 2). In the design-based setting, this irregularity occurred only when the two candidate values for the bounds were equal. Here, the optimization problem is over the larger set $\{V_t(\lambda) : \lambda \in \{0,1\}^3\}$, and the non-standard limiting distributions arises more generally whenever the minimizer and maximizer of $V_t(\lambda)$ are not unique.

THEOREM 3. *Under Assumptions 1, 3 and 4 and standard regularity conditions for estimating parametric models, the vector of six estimated counterfactual means is consistent and asymptotically normal. Consequently, the vector of possible attribution values, $\hat{V}_t = (\hat{V}_t(\lambda))_{\lambda\in\{0,1\}^3}$, is also consistent and asymptotically normal. The estimator for the lower (upper) bound, $\underline{\hat{V}}_t$, is asymptotically normal if the minimum (maximum) of $V_t(\lambda)$ is unique.*[9] *Otherwise, its asymptotic distribution is that of the minimum (maximum) of multiple normal random variables.*

As in the design-based case, Theorem A2 reveals that the estimators for the attribution bounds, $\underline{\hat{V}}_t$ and $\hat{\overline{V}}_t$, have non-standard asymptotic distributions when the optimization problem defining the true bounds (Lemma 2) has a non-unique solution. The estimators are again subject to a conservative bias, and due to the non-differentiability of the min / max operators, standard inferential procedures, such as the bootstrap, can fail to provide correct coverage.

Fortunately, the robust subsampling strategy from Algorithm 1 remains asymptotically valid. To construct confidence intervals in the observational setting, one simply applies the steps of Algorithm 1, with one key modification: we now compute the full-sample and subsample estimates (Steps 1 and 5) using the estimation procedure outlined in Algorithm 2. While this approach is computationally intensive, it correctly accounts for the uncertainty arising from the estimation of the nuisance functions and is valid even when the estimated bounds limiting distribution is non-standard.

4. Empirical Applications

To demonstrate the applicability of our proposed attribution bounds, this section estimates them in two distinct empirical applications. First, I examine a framing experiment

[9] Letting $V_{t,(1)} \leq V_{t,(2)} \leq \ldots \leq V_{t,(8)}$ denote the order statistics for the vector of true values $\{V_t(\lambda)\}_{\lambda\in\{0,1\}^3}$, the condition for a unique minimum is $V_{t,(1)} < V_{t,(2)}$, and for a unique maximum is $V_{t,(8)} > V_{t,(7)}$.

that was particularly interested in understanding the role of mediators (Brader et al. 2008). Here, I find that the jointly-produced effects of both causes are substantial leading the attribution bounds to be wide. Second, I re-analyze an experiment with a factorial design (Gertler et al. 2014). Here, the estimated attribution bounds are tight, reflecting the fact that the experimental design ensured some of the joint-effects are zero. This example demonstrates that in some applications, there can be limited ambiguity in attribution, and what is the case, our attribution bounds reflect that. In each application, the estimated bounds are shown to transparently communicate the scope for strategic selection of attribution rules, which proves minimal in the first case and substantial in the second.

4.1. A Mediation Study

In this section, I consider a setting where the causal pathways are not known a priori and are an important object of interest. I use data from a classic framing experiment by Brader et al. (2008). In their study, participants were shown a news story about immigration, either with a focus on latino or European immigrants; the authors classify the two articles as treatment and control, respectively. I refer to this as the framing treatment. The authors then measured a potential mediator: the individuals level of anxiety. In the language of this paper, the framing treatment is the first cause, and anxiety is the second cause. Finally, the authors measured negative sentiment towards immigrants, the outcome of interest.

Here, sequential unconfoundedness is not guaranteed by design and must be assumed. The framing treatment (A_1) is randomly assigned, but anxiety (A_2) is not. Here, sequential unconfoundedness requires that, conditional on the framing treatment and pre-treatment covariates, anxiety is as-good-as-randomly assigned with respect to potential outcomes (negative sentiment towards immigrants). While this is a strong assumption, it is standard in the mediation literature (Imai et al. 2010a). Under this assumption, I estimate attribution bounds for the framing treatment and anxiety.[10]

To keep the estimation simple and in line with practice, I fit linear models of the expected potential outcomes and potential treatments. However, I use K-fold cross-fitting to ensure that each individual's predicted counterfactuals are not based on their own data.[11] I use the sample of 265 individuals that have complete records.

[10] To fit in our framework, I discretize the authors' continuous measure of anxiety into a binary variable.

[11] In practice, I use 10 folds.

TABLE 1. Attribution Bounds for the Framing Experiment Brader et al. (2008)

	Lower Bound	**Upper Bound**
Framing Treatment	0.037	0.156
	[-0.017, 0.106]	[0.051, 0.229]
Anxiety	0.295	0.413
	[0.197, 0.394]	[0.304, 0.505]
Combined Effect	0.451	
	[0.340, 0.545]	

	Estimate	**95% Confidence Interval**
$\mathbb{E}[D_1(\theta)]$	0.087	[-0.006, 0.176]
$\mathbb{E}[D_2(\theta)]$	0.345	[0.243, 0.437]
$\mathbb{E}[I(\theta)]$	0.068	[0.031, 0.100]
$\mathbb{E}[X_D(\theta)]$	-0.025	[-0.062, 0.018]
$\mathbb{E}[X_I(\theta)]$	-0.024	[-0.058, 0.015]

Notes: Results from a re-analysis of Brader et al. (2008). The first cause is the framing treatment and the second is anxiety. The outcome is standardized negative sentiment towards immigrants. Point estimates in the top table represent the lower and upper attribution bounds for each cause, with 95% confidence intervals for those bounds shown in brackets below. Confidence intervals are from 2000 subsamples of size $b = N^{4/5}$.

Table 1 presents the results. The analysis seeks to explain the causes' combined effect, captuing the impact of both the framing treatment and anxiety.[12]

The combined effect of the intervention is a 0.45 standard deviation increase in negative sentiment. More importantly for the purposes of this paper, the jointly-produced effects are substantial, leading to wide attribution bounds (a range of approximately 0.12 standard deviations). The ambiguity arises from the moderately sized indirect and indirect interaction effects.

The wide range reveals significant attributional ambiguity and highlights the value of this paper's approach. A researcher aiming to downplay the importance of racial cues in the news could strategically select a causal decomposition that yields attribution

[12]Note: The combined effect is distinct from the total effect considered in the mediation literature: $\mathbb{E}[Y(1, A_2(1)) - Y(0, A_2(0))]$ as the combined effect, $\mathbb{E}[Y(A_1, A_2(A_1)) - Y(0, 0))]$, captures the effect of both causes, including any baseline effect of anxiety.

values near or at the lower bound for the framing treatment making it seem statistically insignificant. Conversely, a researcher wishing to emphasize the importance of the framing treatment could select a different causal decomposition yielding attribution values near or at the upper bound, which is statistically significant. By reporting the full bounds, the scope for this kind of strategic behavior is made transparent. Instead of a single, carefully selected point-estimate, the bounds present the full range of conclusions supported by the data across all valid attribution rules.

Moreover, the attribution bounds can still yield decisive conclusions. Notice that the entire confidence interval for anxiety is above zero. This implies that regardless of the chosen attribution rule, the credit assigned to anxiety is positive and statistically significant. One can therefore conclude that anxiety is an important driver of sentiment towards immigrants in this experiment, a conclusion that is robust to the ambiguity of attribution.

4.2. An Experiment with a Factorial Design

The second application uses data from a study by Gertler et al. (2014), which follows up on an experiment run in Jamaica a few decades earlier Grantham-McGregor et al. (1991). The original experiment provided growth-stunted toddlers with either a nutritional supplement, psychosocial stimulation, or both. Each treatment arm in the original experiment had 32 participants, while the control arm had 33 participants. The authors were able to re-interview 105 out of the 129 original participants and found limited evidence for selective attrition.

To run the analysis, I leverage the experiment's factorial design. While the two treatments were assigned independently, I conduct an "as-if" sequential analysis. I impose a hypothetical causal ordering, treating psychosocial stimulation as the first cause (A_1) and nutritional supplementation as the second (A_2). Under this imposed structure, the factorial design's independence ensures two things. First, the effect of the first cause on the second must be zero, meaning the true indirect and indirect interaction effects are zero. The method should recover this. Second, because both treatments were randomized, the identifying assumption, sequential unconfoundedness, holds by design.

Table 2 presents the estimated attribution bounds. Here, the goal is to explain the average combined effect of the realized treatment path, $\mathbb{E}[Y(A_1, A_2(A_1)) - Y(0, 0)]$, which in this context represents the effect on log earnings. The estimated combined effect is a sizeable 0.119 increase, though it is not statistically significant at conventional levels;

TABLE 2. Attribution Bounds for Gertler et al. (2014)

	Lower Bound	Upper Bound
Psychosocial Stimulation	0.172	0.178
	(0.085)	(0.063)
Nutritional Supplement	-0.060	-0.053
	(0.518)	(0.548)
Combined Effect	0.119	
	(0.355)	

	Estimate	P-Value
$\mathbb{E}[D_1(\theta)]$	0.178	(0.107)
$\mathbb{E}[D_2(\theta)]$	-0.053	(0.623)
$\mathbb{E}[I(\theta)]$	-0.003	(0.611)
$\mathbb{E}[X_D(\theta)]$	-0.003	(0.963)
$\mathbb{E}[X_I(\theta)]$	-0.000	(0.981)

Notes: Results from a re-analysis of Gertler et al. (2014). The first cause is psychosocial stimulation, the second cause is nutritional supplementation, and the outcome is log-earnings. P-values from permutation tests are reported in parentheses.

reported p-values are from a permutation test, in line with the original authors.

As the bottom panel of Table 2 shows, the estimated indirect and indirect interaction effects ($\mathbb{E}[I(\theta)]$ and $\mathbb{E}[X_I(\theta)]$) are effectively zero, affirming our expectation from the design. More interestingly, the direct interaction effect ($\mathbb{E}[X_D(\theta)]$) is also negligible. Because the ambiguity in attribution stems from these three jointly-produced effects, their near-zero magnitude results in extremely tight attribution bounds. As a consequence, the problem of strategically selecting a causal decomposition is moot here; the attribution bounds indicate that any causal decomposition would yield nearly identical conclusions. The tightness of the bounds is an important finding worth reporting in and of itself as it provides quantitative evidence that this particular causal attribution is robust and not subject to meaningful debate.

5. Conclusion

This paper addresses a common question in applied research: "How much did each cause contribute to their combined effect?" I argue that the prevailing practice of reporting a single number masks a fundamental conceptual ambiguity rather than resolving it. The paper's central insight is to move from debating which single decomposition is best to formally characterizing the set of all causal decompositions. The analysis establishes, in two complementary ways, that a decomposition is causal if and only if it attributes to each cause its individually-produced effects while allowing any convex split of the jointly-produced effects. This characterization reveals that ambiguity in attribution is an inherent feature of multi-cause systems, stemming entirely from the different ways one might attribute these joint effects. This characterization, in turn, provides a common language for interpreting existing methods, showing that approaches, such as various Shapley values and sequential decompositions, represent different implicit choices for how to handle this ambiguity.

This insight leads directly to the paper's main proposal: embrace ambiguity by reporting attribution bounds. These bounds represent the minimum and maximum contributions attributable to a cause across all causal decompositions. They transparently communicate the degree of ambiguity, collapsing to a point only when no joint effects are present. To ensure this approach is practical, I develop identification and estimation strategies for both experimental and observational settings. As the empirical applications demonstrate, the bounds directly quantify the ambiguity: they are tight when joint effects are minimal, but widen to reveal the ambiguity that a single point estimate would otherwise hide.

Beyond offering a new method, this work also clarifies the proper role of attribution. This reframes its primary application: not as a decision-making tool, but as a robust method for clear communication. When a decision problem can be fully specified and its solution implemented, it should be, as it leads to better decisions. The value of attribution bounds, therefore, lies in distilling the complex causal structure of a system into an honest and transparent summary, which is especially valuable when the ultimate decision problem is vague, if not unknown, or when results must be conveyed to diverse audiences.

By providing a formal language for the ambiguity in attributing the combined effect of multiple causes, this work contributes to and clarifies debates across several literatures. It reframes the search for a single multi-touch attribution model in marketing

as a choice about how to handle joint effects; it offers a causal interpretation of the path-dependency problem in economics; and it bridges the "Effects-of-Causes" and "Causes-of-Effects" literature by focusing on the unique challenge of attributing an average effect.

Ultimately, by replacing the pretense of a single answer with a transparent account of the underlying ambiguity, this work provides a more honest and robust foundation for reasoning about causal attribution.

References

Andrews, Isaiah and Jesse M Shapiro (2021) "A model of scientific communication," *Econometrica*, 89 (5), 2117–2142.

Badinski, Ivan, Amy Finkelstein, Matthew Gentzkow, and Peter Hull (2023) "Geographic variation in healthcare utilization: The role of physicians," technical report, National Bureau of Economic Research.

Bauer, Thomas K and Mathias Sinning (2008) "An extension of the Blinder–Oaxaca decomposition to nonlinear models," *AStA Advances in Statistical Analysis*, 92 (2), 197–206.

Berman, Ron (2018) "Beyond the last touch: Attribution in online advertising," *Marketing Science*, 37 (5), 771–792.

Blinder, Alan S (1973) "Wage discrimination: reduced form and structural estimates," *Journal of Human resources*, 436–455.

Brader, Ted, Nicholas A Valentino, and Elizabeth Suhay (2008) "What triggers public opposition to immigration? Anxiety, group cues, and immigration threat," *American journal of political science*, 52 (4), 959–978.

Budhathoki, Kailash, Lenon Minorics, Patrick Blöbaum, and Dominik Janzing (2022) "Causal structure-based root cause analysis of outliers," in *International Conference on Machine Learning*, 2357–2369, PMLR.

Bugni, Federico A, Ivan A Canay, and Steve McBride (2025) "Decomposition and interpretation of treatment effects in settings with delayed outcomes," *arXiv preprint arXiv:2302.11505*.

Chernozhukov, Victor, Iván Fernández-Val, and Blaise Melly (2013) "Inference on counterfactual distributions," *Econometrica*, 81 (6), 2205–2268.

Dalessandro, Brian, Claudia Perlich, Ori Stitelman, and Foster Provost (2012) "Causally motivated attribution for online advertising," in *Proceedings of the sixth international workshop on data mining for online advertising and internet economy*, 1–9.

DiNardo, John, Nicole M. Fortin, and Thomas Lemieux (1996) "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica*, 64 (5), 1001–1044, `http://www.jstor.org/stable/2171954`.

Du, Ruihuan, Yu Zhong, Harikesh Nair, Bo Cui, and Ruyang Shou (2019) "Causally driven incremental multi touch attribution using a recurrent neural network," *arXiv preprint arXiv:1902.00215*.

Econsultancy and Google Analytics (2012) "Marketing Attribution: Valuing the Customer Journey," white paper, Google Analytics and Econsultancy, `https://services.google.com/fh/files/misc/marketing_attribution_whitepaper.pdf`.

Fairlie, Robert W (2005) "An extension of the Blinder-Oaxaca decomposition technique to logit and probit models," *Journal of economic and social measurement*, 30 (4), 305–316.

Fortin, Nicole, Thomas Lemieux, and Sergio Firpo (2011) "Decomposition methods in economics," in *Handbook of labor economics*, Vol. 4, 1–102: Elsevier.

Frye, Christopher, Colin Rowat, and Ilya Feige (2020) "Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability," *Advances in Neural Information Processing Systems*, 33, 1229–1239.

Gelman, Andrew (2011) "Causality and statistical learning," *American Journal of Sociology*.

Gelman, Andrew and Guido Imbens (2013) "Why ask why? Forward causal inference and reverse causal questions," technical report, National Bureau of Economic Research.

Gertler, Paul, James Heckman, Rodrigo Pinto, Arianna Zanolini, Christel Vermeersch, Susan Walker, Susan M Chang, and Sally Grantham-McGregor (2014) "Labor market returns to an early childhood stimulation intervention in Jamaica," *Science*, 344 (6187), 998–1001.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár (2024) "Contamination bias in linear regressions," *American Economic Review*, 114 (12), 4015–4051.

Grantham-McGregor, Sally M, Christine A Powell, Susan P Walker, and John H Himes (1991) "Nutritional supplementation, psychosocial stimulation, and mental development of stunted children: the Jamaican Study," *The Lancet*, 338 (8758), 1–5.

Greenland, S and JM Robins (1988) "Conceptual problems in the definition and interpretation of attributable fractions," *American journal of epidemiology*, 128 (6), 1185—1197, 10.1093/oxfordjournals.aje.a115073.

Hernán, Miguel A and James M Robins (2020) *Causal Inference: What If*, Boca Raton: Chapman Hall/CRC.

Heskes, Tom, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen (2020) "Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models," *Advances in neural information processing systems*, 33, 4778–4789.

Holland, Paul W (1986) "Statistics and causal inference," *Journal of the American statistical Association*, 81 (396), 945–960.

Imai, Kosuke, Luke Keele, and Dustin Tingley (2010a) "A general approach to causal mediation analysis.," *Psychological methods*, 15 (4), 309.

Imai, Kosuke, Luke Keele, and Teppei Yamamoto (2010b) "Identification, inference and sensitivity analysis for causal mediation effects," *Statistical Science*, 25 (1), 51–71.

Imbens, Guido W. and Joshua D. Angrist (1994) "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62 (2), 467–475, http://www.jstor.org/stable/2951620.

Jung, Yonghan, Shiva Kasiviswanathan, Jin Tian, Dominik Janzing, Patrick Blöbaum, and Elias Bareinboim (2022) "On measuring causal contributions via do-interventions," in *International Conference on Machine Learning*, 10476–10501, PMLR.

Kelly, Joseph, Jon Vaver, and Jim Koehler (2018) "A causal framework for digital attribution."

Kitagawa, Evelyn M (1955) "Components of a difference between two rates," *Journal of the american statistical association*, 50 (272), 1168–1194.

Kitts, Brendan, Liang Wei, Dyng Au, Amanda Powter, and Brian Burdick (2010) "Attribution of conversion events to multi-channel media," in *2010 IEEE International Conference on Data Mining*, 881–886, IEEE.

Li, Hongshuang and PK Kannan (2014) "Attributing conversions in a multichannel online marketing environment: An empirical model and a field experiment," *Journal of Marketing Research*, 51 (1), 40–56.

Li, Hongshuang, PK Kannan, Siva Viswanathan, and Abhishek Pani (2016) "Attribution strategies and return on keyword investment in paid search advertising," *Marketing Science*, 35 (6), 831–848.

Machado, José AF and José Mata (2005) "Counterfactual decomposition of changes in wage distributions using quantile regression," *Journal of applied Econometrics*, 20 (4), 445–465.

Mill, John Stuart (2011 [1843]) *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence, and the Methods of Scientific Investigation*, Cambridge Library Collection - Philosophy: Cambridge University Press.

Oaxaca, Ronald (1973) "Male-female wage differentials in urban labor markets," *International economic review*, 693–709.

Pearl, Judea (1999) "Probabilities of Causation: Three Counterfactual Interpretations and Their Identification," *Synthese*, 93–149.

——— (2009) *Causality*, Cambridge university press.

Politis, Dimitris N and Joseph P Romano (1994) "Large sample confidence regions based on subsamples under minimal assumptions," *The Annals of Statistics*, 2031–2050.

Quantcast (2012) "Beyond Last Touch: Understanding Campaign Effectiveness," `http://info.quantcast.com/rs/quantcast/images/QuantcastWhitePaper_BeyondLastTouch.pdf`.

Robins, James M and Sander Greenland (1992) "Identifiability and exchangeability for direct and indirect effects," *Epidemiology*, 3 (2), 143–155.

Rosenbaum, Paul R (2001) "Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot," *Biometrika*, 88 (1), 219–231.

Shao, Xuhui and Lexin Li (2011) "Data-driven multi-touch attribution models," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 258–264.

Shapley, LS (1952) "A Value for n-Person Games. RAND Paper P-295."

Shorrocks, Anthony (2013) "Decomposition procedures for distributional analysis: a unified framework based on the Shapley value," *The Journal of Economic Inequality*, 11 (1), 99–126.

Singal, Raghav, Omar Besbes, Antoine Desir, Vineet Goyal, and Garud Iyengar (2022) "Shapley Meets Uniform: An Axiomatic Framework for Attribution in Online Advertising," *Management Science*, 68 (10), 7457–7479.

Singal, Raghav, George Michailidis, and Hoiyi Ng (2021) "Flow-based attribution in graphical models: A recursive shapley approach," in *International Conference on Machine Learning*, 9733–9743, PMLR.

VanderWeele, Tyler J (2014) "A unification of mediation and interaction: a 4-way decomposition," *Epidemiology*, 25 (5), 749–761.

VanderWeele, Tyler J and Eric J Tchetgen Tchetgen (2014) "Attributing effects to interactions," *Epidemiology*, 25 (5), 711–722.

Yamamoto, Teppei (2012) "Understanding the past: Statistical analysis of causal attribution," *American Journal of Political Science*, 56 (1), 237–256.

Appendix A. Proofs for Section 2

Here we provide the proofs for Section 2. In order, we prove:

(a) Lemma 1, which establishes the algebraic identity between the total effect and the sum of its five constituent causal components.

(b) Proposition 1, which establishes that on the set of non-defiers, satisfying the first three properties is equivalent to being a causal decomposition.

(c) Theorem 1, which establishes that satisfying the four properties is equivalent to being a causal decomposition.

To aid our discussion, we find it useful to first introduce some notation. First, note that for any binary variables $x_1, x_2 \in \{0, 1\}$ we have:

$$\begin{aligned} y(x_1, x_2) = \; & y(0,0) + [y(1,0) - y(0,0)]x_1 + [y(0,1) - y(0,0)]x_2 \\ & + [y(1,1) - y(0,1) - y(1,0) + y(0,0)]x_1 x_2 \end{aligned}$$

We define the coefficients here as β_1, β_2 and β_{12} respectively:

$$y(x_1, x_2) = y(0,0) + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

Using this notation, we can rewrite the five causal components compactly as follows:

$$\begin{aligned} D_1(\theta) &= \big[y(1,0) - y(0,0)\big]\, a_1 && = a_1 \beta_1 \\ D_2(\theta) &= \big[y(0,1) - y(0,0)\big]\, a_2(0) && = a_2(0)\beta_2 \\ I(\theta) &= \big[y(0,1) - y(0,0)\big]\, (a_2(1) - a_2(0))\, a_1 && = a_1(a_2(1) - a_2(0))\beta_2 \\ X_D(\theta) &= \big[y(1,1) - y(0,1) - (y(1,0) - y(0,0))\big]\, a_2(0) && = a_1 a_2(0)\beta_{12} \\ X_I(\theta) &= \big[y(1,1) - y(0,1) - (y(1,0) - y(0,0))\big]\, (a_2(1) - a_2(0))\, a_1 && = a_1(a_2(1) - a_2(0))\beta_{12} \end{aligned}$$

We now proceed to demonstrate the desired results.

A.1. Proof of Lemma 1

LEMMA 1 (continuing from p. 13). *For type, $\theta = (a_1, a_2(\cdot), y(\cdot)) \in \Theta$, the total effect, $\tau(\theta)$, can always be written as the sum of the following five component effects:*

$$\tau(\theta) = D_1(\theta) + D_2(\theta) + I(\theta) + X_D(\theta) + X_I(\theta)$$

where each component is defined as:

$$
\begin{aligned}
D_1(\theta) &= \big[\, y(1,0) - y(0,0) \big]\, a_1 \\
D_2(\theta) &= \big[\, y(0,1) - y(0,0) \big]\, a_2(0) \\
I(\theta) &= \big[\, y(0,1) - y(0,0) \big]\, (a_2(1) - a_2(0))\, a_1 \\
X_D(\theta) &= \big[\, y(1,1) - y(0,1) - (\, y(1,0) - y(0,0)) \big]\, a_2(0) a_1 \\
X_I(\theta) &= \big[\, y(1,1) - y(0,1) - (\, y(1,0) - y(0,0)) \big]\, (a_2(1) - a_2(0))\, a_1
\end{aligned}
$$

PROOF. We prove the identity by direct algebraic manipulation. We begin by rewriting $\tau(\theta) = y(a_1, a_2(a_1)) - y(0,0)$ in terms of the β's defined in the section's preamble (Section A):

$$
\begin{aligned}
\tau(\theta) &= y(a_1, a_2(a_1)) - y(0,0) \\
&= \beta_1 a_1 + \beta_2 a_2(a_1) + \beta_{12} a_1 a_2(a_1).
\end{aligned}
$$

Moreover, we can always expand the potential treatments as follows, for any $x_1 \in \{0,1\}$:

$$
a_2(\tilde{x}_1) = a_2(0) + [a_2(1) - a_2(0)]\tilde{x}_1
$$

Applying this identity letting $x_1 = a_1$ gives us the following:

$$
\begin{aligned}
\tau(\theta) &= \beta_1 a_1 + \beta_2 a_2(0) + \beta_2 (a_2(1) - a_2(0)) a_1 + \beta_{12} a_1 a_2(0) + \beta_{12} (a_2(1) - a_2(0)) a_1 \\
&= D_1(\theta) + D_2(\theta) + I(\theta) + X_D(\theta) + X_I(\theta)
\end{aligned}
$$

where the last step follows by definition (see Section A preamble). This completes the proof. □

A.2. Proof of Proposition 1

PROPOSITION 1 (continuing from p. 17). *For a decomposition, $\psi \in \Psi$, the following two conditions are equivalent on the set of non-defiers, $\{(a_1, a_2(\cdot), y(\cdot)) : a_2(1) \geq a_2(0)\}$:*

(i) *ψ satisfies Linearity, No Credit for No Effect, and No Blame for No Harm (Properties 1, 2, and 3)*

(ii) *ψ is a causal decomposition, $\psi \in \Psi^c$*

PROOF. We prove both directions of the equivalence starting with "Properties ⇐ Causal

Decomposition" and then proceeding to "Properties ⇒ Causal Decomposition". Note that all statements and types in this proof are restricted to the set of non-defiers $\theta \in \Theta_{ND} = \{(a_1, a_2(\cdot), y(\cdot)) : a_2(1) \geq a_2(0)\}$.

Properties ⇐ Causal Decomposition. Assume that for any $\theta \in \Theta_{ND}$, $\psi(\theta) = \psi^c(\theta)$ for some fixed causal decomposition $\psi^c \in \Psi^c$. We need to show that ψ satisfies the three properties for any type in Θ_{ND}.

(a) *Linearity:* A causal decomposition ψ^c is defined as a linear combination of the five components D_1, D_2, I, X_D, X_I. Each of these components is a linear function of the potential outcome function $y(\cdot)$. For example, $D_1(\theta; y + \tilde{y}) = [(y(1,0) + \tilde{y}(1,0)) - (y(0,0) + \tilde{y}(0,0))]a_1 = D_1(\theta; y) + D_1(\theta; \tilde{y})$. Since ψ^c is a sum of such terms, it satisfies additivity. Similarly, scaling $y(\cdot)$ by c scales each component by c, so ψ^c satisfies homogeneity. Thus, ψ^c satisfies Linearity.

(b) *No Credit for No Effect:*

(i) For Cause 1: Recall $\psi_1^c = D_1 + \lambda_1 I + \lambda_2 X_D + \lambda_3 X_I$. If $a_1 = 0$, all four of these components are zero, so $\psi_1^c(\theta) = 0$. If $m_1(\cdot; \theta) = 0$, then $y(1,0) = y(0,0)$ and $y(1,1) = y(0,1)$. This implies $D_1(\theta) = 0$ and the interaction term $[y(1,1) - y(0,1) - (y(1,0) - y(0,0))]$ is also zero, which in turn means $X_D(\theta) = 0$ and $X_I(\theta) = 0$. The property as stated also requires $m_{1\to 2}(\cdot; \theta) = 0$. This means $y(b_1, a_2(1)) - y(b_1, a_2(0)) = 0$. If $a_2(1) \neq a_2(0)$, this requires $y(b_1, 1) = y(b_1, 0)$ for $b_1 \in \{0, 1\}$, which implies $m_2(\cdot; \theta) = 0$. If $m_2(0; \theta) = 0$, then $y(0,1) = y(0,0)$, which implies $I(\theta) = 0$. Thus, $\psi_1^c(\theta) = 0$.

(ii) For Cause 2: Recall $\psi_2^c = D_2 + (1-\lambda_1)I + (1-\lambda_2)X_D + (1-\lambda_3)X_I$. If $a_2(0) = a_2(1) = 0$, then $D_2 = I = X_D = X_I = 0$, so $\psi_2^c(\theta) = 0$. If $m_2(\cdot; \theta) = 0$, then $y(0,1) = y(0,0)$ and $y(1,1) = y(1,0)$. This implies $D_2(\theta) = 0$ and $I(\theta) = 0$. The interaction term becomes $[y(1,0) - y(0,0) - (y(1,0) - y(0,0))] = 0$, which means $X_D(\theta) = 0$ and $X_I(\theta) = 0$. Thus, $\psi_2^c(\theta) = 0$.

(c) *No Blame for No Harm:* For non-defiers, $a_2(1) \geq a_2(0)$, so $a_2(1) - a_2(0) \geq 0$. The conditions for Cause 1 imply $m_1(0) \geq 0$, $m_2(0) \geq 0$ (from $m_{1\to 2}$), and the interaction term is non-negative. Each component D_1, I, X_D, X_I is thus a product of non-negative terms, making them non-negative. Since $\lambda_i \in [0, 1]$, ψ_1^c is a sum of non-negative terms and must be non-negative. A symmetric argument holds for ψ_2^c.

$$\psi(\theta') = \psi^c(\theta') = \psi^c(\theta) + \psi^c(\tilde{\theta}) = \psi(\theta) + \psi(\tilde{\theta}).$$

The same logic of inheritance applies directly to Homogeneity, No Credit for No Effect, and No Blame for No Harm.

***Properties* ⇒ *Causal Decomposition*.** Assume ψ satisfies the three properties for all $\theta \in \Theta_{ND}$. We show that ψ must take the form of a causal decomposition.

(a) *Linearity.* The Linearity property implies that for any given $\theta = (a_1, a_2(\cdot), y(\cdot)) \in \Theta_{ND}$, $\psi_1(\theta)$ is a linear functional of the potential outcome function $y(\cdot)$. We can therefore express it as a linear combination of a basis for the space of outcome functions. A convenient basis is given by the constant outcome $y(0,0)$ and the principal effect terms defined in the section preamble as β's (see Section A):

$$\psi_1(\theta) = f_0(\theta_A)\, y(0,0) + f_1(\theta_A)\beta_1 + f_2(\theta_A)\beta_2 + f_{12}(\theta_A)\beta_{12}$$

where the coefficients f_j depend only on the treatment path characteristics $\theta_A = (a_1, a_2(0), a_2(1))$.

First, note that if all effect terms to zero ($\beta_1 = \beta_2 = \beta_{12} = 0$), then all marginal effects are zero, so by No Credit for No Effect (Property 2), $\psi_1(\theta)$ must be zero. The formula gives $\psi_1(\theta) = f_0(\theta_A)\, y(0,0)$. For this to be zero for any $y(0,0)$, we must have $f_0(\theta_A) = 0$. Our representation simplifies to:

$$\psi_1(\theta) = f_1(\theta_A)\beta_1 + f_2(\theta_A)\beta_2 + f_{12}(\theta_A)\beta_{12}$$

(b) *Determine* $f_1(\theta_A)$. Consider a type where only $\beta_1 \neq 0$. Then $\psi_1(\theta) = f_1(\theta_A)\beta_1$. In this case, the second cause has no marginal effects ($m_2(\cdot) = 0$), so by Property 2(ii), $\psi_2(\theta) = 0$. The total effect is $\tau(\theta) = y(a_1, a_2(a_1)) - y(0,0) = a_1\beta_1$. The adding-up constraint $\psi_1 + \psi_2 = \tau$ implies $\psi_1(\theta) = a_1\beta_1$. Therefore, $f_1(\theta_A)\beta_1 = a_1\beta_1$, which means $\boldsymbol{f_1(\theta_A) = a_1}$.

(c) *Determine* $f_2(\theta_A)$. Consider a type where only $\beta_2 \neq 0$. Then $\psi_1(\theta) = f_2(\theta_A)\beta_2$. In this scenario, cause 1 has no direct marginal effect ($m_1(\cdot) = 0$). By Property 2(i), if cause 1 is also inactive ($a_1 = 0$) or has no indirect effect ($m_{1\to 2}(\cdot) = 0$), it gets no credit.

- If $a_1 = 0$, then $\psi_1(\theta) = 0$, so $f_2(0, a_2(0), a_2(1)) = 0$.

- If $a_2(1) = a_2(0)$ (i.e., always-taker or never-taker), then $m_{1\to 2}(\cdot) = \beta_2(a_2(1) - a_2(0)) = 0$. So $\psi_1(\theta) = 0$, which implies $f_2(a_1, x, x) = 0$ for $x \in \{0, 1\}$.

The only function f_2 of three binary variables $(a_1, a_2(0), a_2(1))$ that is zero when $a_1 = 0$ or when $a_2(0) = a_2(1)$ must be proportional to $a_1(a_2(1) - a_2(0))$. So we can write $f_2(\theta_A) = \lambda_1 a_1(a_2(1) - a_2(0))$ for some constant λ_1. Using No Blame for No Harm on a complier type with $\beta_2 > 0$ confirms $\lambda_1 \in [0, 1]$.

(d) *Determine* $f_{12}(\theta_A)$. Consider a type where only $\beta_{12} \neq 0$. Then $\psi_1(\theta) = f_{12}(\theta_A)\beta_{12}$.

- If $a_1 = 0$, Property 2(i) gives $\psi_1(\theta) = 0$, so $f_{12}(0, a_2(0), a_2(1)) = 0$. This implies f_{12} must be proportional to a_1.

So, $f_{12}(\theta_A)$ must be of the form $a_1 \cdot g(a_2(0), a_2(1))$ for some function g defined on two binary inputs. We don't care about $g(1, 0)$ because we're not discussing defiers, and we don't care about $g(0, 0)$, because there's no credit to be distributed. The two cases of interest are when $g(0, 1)$ and $g(1, 1)$. In both cases, so long as $\beta_{12} > 0$ No Blame for No Harm ensures that the coefficients are non-negative, so we get:

$$f_{12}(\theta_A) = a_1 \left(\lambda_2 a_2(0) + \lambda_3(a_2(1) - a_2(0))\right)$$

for $\lambda_2, \lambda_3 \geq 0$

(e) *Putting it all together.* Combining these findings, we have:

$$\begin{aligned}\psi_1(\theta) &= a_1\beta_1 + \left[\lambda_1 a_1(a_2(1) - a_2(0))\right]\beta_2 + \left[a_1\left(\lambda_2 a_2(0) + \lambda_3(a_2(1) - a_2(0))\right)\right]\beta_{12} \\ &= a_1\beta_1 + \lambda_1\beta_2(a_2(1) - a_2(0))a_1 + \lambda_2\beta_{12}a_2(0)a_1 + \lambda_3\beta_{12}(a_2(1) - a_2(0))a_1\end{aligned}$$

Recalling the definitions of the five components in terms of the βs from the section's preamble (see Section A), we can rewrite our expression for $\psi_1(\theta)$ by substituting these back in:

$$\psi_1(\theta) = D_1(\theta) + \lambda_1 I(\theta) + \lambda_2 X_D(\theta) + \lambda_3 X_I(\theta).$$

This is precisely the formula for a causal decomposition, with sharing parameters $(\lambda_1, \lambda_2, \lambda_3) \in [0, 1]^3$. By the adding-up constraint, $\psi_2(\theta) = \tau(\theta) - \psi_1(\theta)$. Substituting the full expression for $\tau(\theta)$ from Lemma 1 and the derived expression for $\psi_1(\theta)$, one

can write $\psi_2(\theta)$ in the desired form:

$$\begin{aligned}\psi_2(\theta) &= \Big(D_1(\theta) + D_2(\theta) + I(\theta) + X_D(\theta) + X_I(\theta)\Big) \\ &\qquad - \Big(D_1(\theta) + \lambda_1 I(\theta) + \lambda_2 X_D(\theta) + \lambda_3 X_I(\theta)\Big) \\ &= D_2(\theta) + (1-\lambda_1) I(\theta) + (1-\lambda_2) X_D(\theta) + (1-\lambda_3) X_I(\theta).\end{aligned}$$

This completes the proof. □

A.3. Proof of Theorem 1

THEOREM 1 (continuing from p. 18). *A decomposition, $\psi \in \Psi$, satisfies Linearity, No Credit for No Effect, No Blame for No Harm, and Symmetry of Causal Changes if and only if it is a causal decomposition, $\psi \in \Psi^c$.*

PROOF. We prove both directions of the equivalence starting with "Properties $\Leftarrow$ Causal Decomposition" and then proceeding to "Properties $\Rightarrow$ Causal Decomposition".

Properties $\Leftarrow$ Causal Decomposition. Consider a causal decomposition $\psi^c \in \Psi$ with fixed sharing parameters $\lambda \in [0,1]^3$. We have already shown in the proof of Proposition 1 that any causal decomposition satisfies Linearity, No Credit for No Effect, and No Blame for No Harm for all non-defier types $\theta \in \Theta_{ND}$. The same arguments for Linearity and No Credit for No Effect apply for the defiers as well. However, we do need to prove that No Blame for No Harm applies to defiers, and of course we need to prove that Symmetry of Causal Changes is satisfied.

(a) We now show a causal decomposition satisfies No Blame for No Harm on the defiers, i.e., $a_2(1) < a_2(0)$. We need to check two conditions

- If $\beta_1 \geq 0$, $\beta_2 \leq 0$ and $\beta_{12} = 0$, we need $\psi_1^c \geq 0$. This is satisfied, because the first cause gets credit for $D_1(\theta) + \lambda_1 I(\theta) = \beta_1 a_1 + \lambda_1 \beta_2 (a_2(a_1) - a_2(0)) \geq 0$, where the second term is non-negative because $\beta_2 \leq 0$ and $a_2(1) < a_2(0)$.
- If $\beta_2 = \beta_{12} = 0$, then $\psi_2^c \geq 0$. This is satisfied because the second cause gets no credit, $\psi_2^c(\theta) = 0$.

(b) We now show it also satisfies Symmetry of Causal Changes. We must verify:

$$\psi^c(a_1, a_2^C, y) - \psi^c(a_1, a_2^{NT}, y) = \psi^c(a_1, a_2^{AT}, y) - \psi^c(a_1, a_2^D, y).$$

Let's analyze the first component, ψ_1^c. The proof for ψ_2^c follows from the adding-up constraint. Noting that the form of ψ_1^c is $D_1 + \lambda_1 I + \lambda_2 X_D + \lambda_3 X_I$, we rewrite each term:

- Complier (θ^C): $a_2^C(0) = 0$, $a_2^C(1) = 1$. $\psi_1^c(\theta^C) = D_1 a_1 + \lambda_1 \beta_2 a_1 + \lambda_3 \beta_{12} a_1$.
- Never-Taker (θ^{NT}): $a_2^{NT}(0) = 0$, $a_2^{NT}(1) = 0$. All jointly-produced effects are zero. $\psi_1^c(\theta^{NT}) = D_1 a_1$.
- Always-Taker (θ^{AT}): $a_2^{AT}(0) = 1$, $a_2^{AT}(1) = 1$. $\psi_1^c(\theta^{AT}) = D_1 a_1 + \lambda_2 \beta_{12} a_2(0) a_1$.
- Defier (θ^D): $a_2^D(0) = 1$, $a_2^D(1) = 0$. $\psi_1^c(\theta^D) = D_1 a_1 - \lambda_1 \beta_2 a_1 + \lambda_2 a_2(0) a_1 - \lambda_3 a_1$.

where we plug in known forms for the five causal components. Combining terms, we see that the compliers minus the never-takers give us $\lambda_1 \beta_2 a_1 + \lambda_3 \beta_{12} a_1$, while the always-takers minus the defiers give us $\lambda_1 \beta_2 a_1 + \lambda_3 a_1$. Thus, the equality is satisfied and Symmetry of Causal Changes is satisfied.

Properties $\Rightarrow$ Causal Decomposition. Consider a decomposition $\psi \in \Psi$ that satisfies the four properties.

(a) Behavior on Non-Defiers: Proposition 1 states that because ψ satisfies the first three properties, its behavior on the set of non-defiers $\Theta_{ND} = \{\theta : a_2(1) \geq a_2(0)\}$ must be that of a causal decomposition. This means there exists a unique vector of sharing parameters $\lambda = (\lambda_1, \lambda_2, \lambda_3) \in [0, 1]^3$ such that:

$$\psi(\theta) = \psi^c(\theta; \lambda) \quad \text{for all } \theta \in \Theta_{ND}.$$

(b) Bridge to Defiers using Symmetry of Causal Changes: We now show that ψ must also conform to $\psi^c(\cdot; \lambda)$ for defier types. A generic defier type is $\theta_D = (a_1, a_2^D, y)$, where $a_2^D(0) = 1$, $a_2^D(1) = 0$. The fourth property provides an equation that connects defiers to non-defiers. We rearrange it to solve for the decomposition of a defier type, $\psi(\theta_D)$:

$$\psi(\theta_D) = \psi^c(\theta_{AT}; \lambda) - \Big[\psi^c(\theta_C; \lambda) - \psi^c(\theta_{NT}; \lambda)\Big]$$

where we leverage the fact that an always-taker (θ_{AT}), complier (θ_C), and never-taker (θ_{NT}), are all non-defiers and thus can be written in terms of a causal decomposition; note, here that each type only differs in the potential treatments, they each have the arbitrary first treatment, a_1, and potential outcomes, $y(\cdot)$. The final step is to plug-in the known form for a causal decomposition and derive ψ's behavior on the defiers,

starting with ψ_1:

$$\begin{aligned}\psi_1(\theta_D) &= D_1(\theta_{AT}) + \lambda_2 X_D(\theta_{AT}) - \left[D_1(\theta_C) + \lambda_1 I(\theta_C) + \lambda_3 X_I(\theta_C) - D_1(\theta_{NT})\right] \\ &= \beta_1 a_1 + \lambda_2 \beta_{12} a_1 - \left[\beta_1 a_1 + \lambda_1 \beta_2 a_1 + \lambda_3 \beta_{12} a_1 - \beta_1 a_1\right] \\ &= \beta_1 a_1 - \lambda_1 \beta_2 a_1 + \lambda_2 \beta_{12} a_1 - \lambda_3 \beta_{12} a_1 \\ &= D_1(\theta_D) + \lambda_1 I(\theta_D) + \lambda_2 X_D(\theta_D) + \lambda_3 X_I(\theta_D)\end{aligned}$$

By the adding-up constraint, $\psi_2(\theta) = \tau(\theta) - \psi_1(\theta)$. Substituting the full expression for $\tau(\theta)$ from Lemma 1 and the derived expression for $\psi_1(\theta)$, one can write $\psi_2(\theta)$ in the desired form:

$$\begin{aligned}\psi_2(\theta) &= \Big(D_1(\theta) + D_2(\theta) + I(\theta) + X_D(\theta) + X_I(\theta)\Big) \\ &\qquad - \Big(D_1(\theta) + \lambda_1 I(\theta) + \lambda_2 X_D(\theta) + \lambda_3 X_I(\theta)\Big) \\ &= D_2(\theta) + (1 - \lambda_1) I(\theta) + (1 - \lambda_2) X_D(\theta) + (1 - \lambda_3) X_I(\theta).\end{aligned}$$

This shows that the behavior of ψ on defier types is uniquely determined by its behavior on non-defier types, and it must conform to the same causal decomposition rule $\psi^c(\cdot; \lambda)$.

Since $\psi(\theta) = \psi^c(\theta; \lambda)$ holds for both non-defiers (by Proposition 1) and defiers (by the Symmetry property), it holds for all types $\theta \in \Theta$. Therefore, ψ is a causal decomposition. □

Appendix B. Proofs for Section 3

B.1. Proof for Lemma 2

LEMMA 2 (continuing from p. 21). *For $t \in \{1, 2\}$, the attribution bounds, $[\underline{V}_t, \overline{V}_t]$, are the minimum and maximum values of the following function evaluated over the set $\lambda = (\lambda_1, \lambda_2, \lambda_3) \in [0, 1]^3$:*

$$
\begin{aligned}
V_t(\lambda) = {} & \mathbb{E}[Y(0,0)] \times (\lambda_2 - 1) && (i)\\
& + \mathbb{E}[Y(A_1(1), 0)] \times (\mathbf{1}\{t = 1\} - \lambda_2) && (ii)\\
& + \mathbb{E}[Y(0, A_2(0,1))] \times (\mathbf{1}\{t = 2\} - \lambda_1 - \lambda_2 + \lambda_3) && (iii)\\
& + \mathbb{E}[Y(0, A_2(A_1(1), 1))] \times (\lambda_1 - \lambda_3) && (iv)\\
& + \mathbb{E}[Y(A_1(1), A_2(0,1))] \times (\lambda_2 - \lambda_3) && (v)\\
& + \mathbb{E}[Y(A_1(1), A_2(A_1(1), 1))] \times \lambda_3 && (vi)
\end{aligned}
$$

PROOF. By definition, the bounds are the extrema of $V_t(\lambda) = \mathbb{E}[\psi_t(\theta; \lambda)]$ over $\lambda \in [0, 1]^3$. We first derive the expression for $V_1(\lambda)$.

From Definition 4, $V_1(\lambda) = \mathbb{E}[D_1(\theta) + \lambda_1 I(\theta) + \lambda_2 X_D(\theta) + \lambda_3 X_I(\theta)]$. We can express the expectation of each component in terms of expectations of potential outcomes, where some outcomes are indexed by potential treatments:

$$
\begin{aligned}
\mathbb{E}[D_1(\theta)] &= \mathbb{E}[(Y(1,0) - Y(0,0))A_1(1)] = \mathbb{E}[Y(A_1(1), 0) - Y(0,0)]\\
\mathbb{E}[D_2(\theta)] &= \mathbb{E}[(Y(0,1) - Y(0,0))A_2(0,1)] = \mathbb{E}[Y(0, A_2(0,1)) - Y(0,0)]\\
\mathbb{E}[I(\theta)] &= \mathbb{E}[(Y(0,1) - Y(0,0))(A_2(1,1) - A_2(0,1))A_1(1)]\\
&= \mathbb{E}[Y(0, A_2(A_1(1),1)) - Y(0, A_2(0,1))]\\
\mathbb{E}[X_D(\theta)] &= \mathbb{E}[(Y(1,1) - Y(0,1) - Y(1,0) + Y(0,0))A_1(1)A_2(0,1)]\\
&= \mathbb{E}[Y(A_1(1), A_2(0,1))] - \mathbb{E}[Y(0, A_2(0,1))] - \mathbb{E}[Y(A_1(1), 0)] + \mathbb{E}[Y(0,0)]\\
\mathbb{E}[X_I(\theta)] &= \mathbb{E}[(Y(1,1) - Y(0,1) - Y(1,0) + Y(0,0))A_1(1)(A_2(1,1) - A_2(0,1))]\\
&= \mathbb{E}[Y(A_1(1), A_2(A_1(1),1))] - \mathbb{E}[Y(0, A_2(A_1(1),1))] - \mathbb{E}[Y(A_1(1), A_2(0,1))]\\
&\quad + \mathbb{E}[Y(0, A_2(0,1))]
\end{aligned}
$$

Substituting these expressions back into the formula for $V_1(\lambda)$:

$$\begin{aligned}
V_1(\lambda) = &\big(\mathbb{E}[Y(A_1(1),0)] - \mathbb{E}[Y(0,0)]\big) \\
&+ \lambda_1\big(\mathbb{E}[Y(0,A_2(A_1(1),1))] - \mathbb{E}[Y(0,A_2(0,1))]\big) \\
&+ \lambda_2\big(\mathbb{E}[Y(A_1(1),A_2(0,1))] - \mathbb{E}[Y(0,A_2(0,1))] - \mathbb{E}[Y(A_1(1),0)] + \mathbb{E}[Y(0,0)]\big) \\
&+ \lambda_3\big(\mathbb{E}[Y(A_1(1),A_2(A_1(1),1))] - \mathbb{E}[Y(0,A_2(A_1(1),1))] - \mathbb{E}[Y(A_1(1),A_2(0,1))] \\
&\qquad + \mathbb{E}[Y(0,A_2(0,1))]\big)
\end{aligned}$$

We now collect terms based on the six counterfactual expectations:

$$\begin{aligned}
V_1(\lambda) = &\mathbb{E}[Y(0,0)](-1+\lambda_2) \\
&+ \mathbb{E}[Y(A_1(1),0)](1-\lambda_2) \\
&+ \mathbb{E}[Y(0,A_2(0,1))](-\lambda_1-\lambda_2+\lambda_3) \\
&+ \mathbb{E}[Y(0,A_2(A_1(1),1))](\lambda_1-\lambda_3) \\
&+ \mathbb{E}[Y(A_1(1),A_2(0,1))](\lambda_2-\lambda_3) \\
&+ \mathbb{E}[Y(A_1(1),A_2(A_1(1),1))](\lambda_3)
\end{aligned}$$

This provides the expression for $V_1(\lambda)$. To find the expression for $V_2(\lambda)$, note that $V_2(\lambda) = \mathbb{E}[D_2] + (1-\lambda_1)\mathbb{E}[I] + (1-\lambda_2)\mathbb{E}[X_D] + (1-\lambda_3)\mathbb{E}[X_I]$. This shows that $V_2(\lambda)$ has the same functional form as $V_1(\lambda)$, but with different constant terms corresponding to the direct effects. To find an expression for $V_2(\lambda)$ then, we simply need to add $\mathbb{E}[D_2(\theta) - D_1(\theta)]$ to our expression for $V_1(\lambda)$. This gives us a general expression for $V_t(\lambda)$ using an indicator function $\mathbf{1}\{t=1\}$:

$$\begin{aligned}
V_t(\lambda) = &(\lambda_2-1)\mathbb{E}[Y(0,0)] \\
&+ (\mathbf{1}\{t=1\}-\lambda_2)\mathbb{E}[Y(A_1(1),0)] \\
&+ (\mathbf{1}\{t=2\}+\lambda_3-\lambda_1-\lambda_2)\mathbb{E}[Y(0,A_2(0,1))] \\
&+ (\lambda_1-\lambda_3)\mathbb{E}[Y(0,A_2(A_1(1),1))] \\
&+ (\lambda_2-\lambda_3)\mathbb{E}[Y(A_1(1),A_2(0,1))] \\
&+ \lambda_3\mathbb{E}[Y(A_1(1),A_2(A_1(1),1))]
\end{aligned}$$

The attribution bounds $[\underline{V}_t, \overline{V}_t]$ are the minimum and maximum of this linear function over the cube $\lambda \in [0,1]^3$. This completes the proof. □

B.2. Proof for Proposition 2

PROPOSITION 2 (continuing from p. 23). *Suppose Assumption 1 holds. There is no experimental design $P \in \mathcal{E}$ that point-identifies the attribution bounds.*

PROOF. This proof exploits the cross-world nature of terms (iv) and (v) in Lemma 2 to show that, for any fixed design $P \in \mathcal{E}$, there exist two causal models that induce the same distribution of observables under that design, but yield different attribution bounds.

By definition of $\mathcal{E}$, we have $\left(F, A^s\right) \perp \left(\{Y(x)\}_x, A_1(1), \{A_2(x_1, 1))\}_{x_1}\right)$ and $P(F = (1,1)) < 1$. We construct two data-generating processes, Q and $\tilde{Q}$, both satisfying Assumption 1 and the design P, such that the joint law of observables (Y, A, F, A^s) is the same under Q and $\tilde{Q}$, but the attribution bounds differ.

Let $\theta_{b_1 b_2}$ describe an individual with $A_1(1) = 1$, $A_2(0,1) = 0$, $A_2(1,1) = b_2$, and define their potential outcomes as follows:

$$Y(0,0) = Y(1,0) = Y(1,1) = 0$$
$$Y(0,1) = b_1$$

This describe four types of individuals depending on the values of $b_1, b_2 \in \{0,1\}$. Each individual only varies in $A_2(1,1)$ and $Y(0,1)$, otherwise they are all identical.

Now consider two data generating processes, B and C, denoted by $\mathbb{P}_B$ and $\mathbb{P}_C$, respectively:

$$\mathbb{P}_B(\theta_{b_1 b_2}) = 1/4$$
$$\mathbb{P}_C(\theta_{b_1 b_2}) = \begin{cases} 1/4 + \gamma & \text{if } b_1 = b_2 \\ 1/4 - \gamma & \text{else} \end{cases}$$

for some fixed $\gamma \in (0, 1/4)$. Now, note that these two data-generating processes are observationally equivalent. To show observational equivalence it suffices to show they follow the same distribution within each experimental arm because:

$$\mathbb{P}_B(Y, A_1, A_2, F, A^s) = \mathbb{P}_C(Y, A_1, A_2, F, A^s)$$
$$\iff \mathbb{P}_B(Y, A_1, A_2 | F, A^s) P(F, A^s) = \mathbb{P}_C(Y, A_1, A_2 | F, A^s) P(F, A^s)$$

and as this needs to hold for any experiment, $P \in \mathcal{E}$, we need $\mathbb{P}_B(Y, A_1, A_2 | F, A^s) = \mathbb{P}_C(Y, A_1, A_2 | F, A^s)$.

So, consider arbitrary $A^s = a^s \in \mathbb{A}$, and let's consider each $F = f \in \mathbb{F}$ in turn:

- If $F = (0, 0)$, we observe $A_1 = a_1^s$, $A_2 = a_2^s$, and $Y = Y(a_1^s, a_2^s)$. The only potential source of variation across DGPs is the outcome when $a^s = (0, 1)$, but in both cases, the probability of $Y = 1$ and $Y = 0$ is the same: 1/2.

$$\begin{aligned} \mathbb{P}_B(Y = 1) = \mathbb{P}_B(\theta_{10} + \theta_{11}) &= \frac{1}{4} + \frac{1}{4} &&= \frac{1}{2} \\ \mathbb{P}_C(Y = 1) = \mathbb{P}_C(\theta_{10} + \theta_{11}) &= \left(\frac{1}{4} - \gamma\right) + \left(\frac{1}{4} + \gamma\right) &&= \frac{1}{2} \end{aligned}$$

- If $F = (1, 0)$, we observe $A_1 = A_1(1) = 1$, $A_2 = a_2^s$, and $Y = Y(1, a_2^s) = 0$. Here, there is no variation, so they are observationally the same.

- If $F = (0, 1)$, we observe $A_1 = a_1^s$, $A_2 = A_2(a_1^s, 1)$, and $Y = Y(a_1^s, A_2(a_1^s, 1))$. There are two cases to consider here.

  - First, if $a_1^s = 0$, then $A_2 = 0$ and $Y = Y(0, 0) = 0$, so again, there is no variation, they are observationally the same.
  - Second, if $a_1^s = 1$, then $A_2 = A_2(1, 1)$ and $Y = Y(1, A_2(1, 1)) = 0$. Here, the only potential source of variation is the second treatment, but in both cases, the probability of $A_2 = 1$ and $A_2 = 0$ is the same: 1/2:

$$\begin{aligned} \mathbb{P}_B(A_2 = 1) = \mathbb{P}_B(\theta_{01} + \theta_{11}) &= \frac{1}{4} + \frac{1}{4} &&= \frac{1}{2} \\ \mathbb{P}_C(A_2 = 1) = \mathbb{P}_C(\theta_{01} + \theta_{11}) &= \left(\frac{1}{4} - \gamma\right) + \left(\frac{1}{4} + \gamma\right) &&= \frac{1}{2} \end{aligned}$$

- If $F = (1, 1)$, we observe $A_1 = A_1(1) = 1$, $A_2 = A_2(1, 1)$, and $Y = Y(1, A_2(1, 1)) = 0$. Here, the only potential source of variation is the second treatment, but as in the last case, the probability of $A_2 = 1$ and $A_2 = 0$ is the same in both DGPs: 1/2:

Thus, the two distributions are observationally equivalent, and thus which DGP we are in is not identified.

What remains to be shown is that the attribution bounds differ across these two DGPs. Recall, the representation of the bounds derived in Lemma 2 and note that the only term that's potentially non-zero here, is the fourth one, i.e.:

$$V_t(\lambda) = (\lambda_1 - \lambda_3)\mathbb{E}[Y(0, A_2(A_1(1), 1))]$$

Now, note that this expectation differs across the two DGPs:

$$\mathbb{E}_B[Y(0, A_2(A_1(1), 1))] = \mathbb{E}_B[Y(0,1)A_2(1,1)] = \mathbb{P}_B(\theta_{11}) = \frac{1}{4}$$
$$\mathbb{E}_C[Y(0, A_2(A_1(1), 1))] = \mathbb{E}_C[Y(0,1)A_2(1,1)] = \mathbb{P}_C(\theta_{11}) = \frac{1}{4} + \gamma$$

Plugging these expectation into our representation of $V_t(\lambda)$ and maximizing/minimizing, gives us two different sets of bounds. Under $\mathbb{P}_B$, the attribution bounds are $[-1/4\,,\ 1/4]$, while under $\mathbb{P}_C$, they are $[-1/4 - \gamma\,,\ 1/4 + \gamma]$. As $\gamma \neq 0$, the attribution bounds in both DGPs are not the same.

As there are two distributions that are observationally equivalent yet yield distinct attribution bounds, we conclude that the attribution bounds are not identified. □

B.3. Proof for Proposition 3

PROPOSITION 3 (continuing from p. 24). *Under Assumptions 1 and 2, an experimental design, $P \in \mathcal{E}$, point identifies the attribution bounds if and only if it is a partial eligibility design, $P \in \mathcal{E}^{PE}$.*

PROOF. The proof proceeds in three main steps. First, we show that Assumption 2 simplifies the identification problem. Second, we prove sufficiency: that a partial eligibility design ensures the attribution bounds are identified ("⇒"). Third, we prove necessity: that the attribution bounds are identified only if the experimental design is a partial eligibility design.

Step 1. Recall that for the first cause, each causal decomposition takes the following form per Definition 4:

$$\psi_1(\theta) = D_1(\theta) + \lambda_1 I(\theta) + \lambda_2 X_D(\theta) + \lambda_3 X_I(\theta)$$

Note that, given this form if $\mathbb{E}[I(\theta)]$, $\mathbb{E}[X_D(\theta)]$ and $\mathbb{E}[X_I(\theta)]$ all have the same sign, the credit assigned to cause one is either maximized or minimized by setting $\lambda_j = 1$ or $\lambda_j = 0$ for $j = 1, 2, 3$.

Taking this observation to our representation of the bounds from Lemma 2, we note

that the bounds are now extrema of the following function $\tilde{V}_t(\lambda_1)$ for $\lambda_1 \in \{0, 1\}$:

$$\begin{aligned}\tilde{V}_t(\lambda_1) = (\lambda_1 - 1)&\mathbb{E}[Y(0,0)] \\ &+ (\mathbf{1}\{t = 1\} - \lambda_1)\mathbb{E}[Y(A_1(1), 0)] \\ &+ (\mathbf{1}\{t = 2\} - \lambda_1)\mathbb{E}[Y(0, A_2(0, 1))] \\ &+ \lambda_1\mathbb{E}[Y(A_1(1), A_2(A_1(1), 1))]\end{aligned}$$

Now, the problem of point-identifying the bounds boils down to point-identifying these four terms.

Step 2. Sufficiency (If $P \in \mathcal{E}^{PE}$, then the attribution bounds are identified)

Consider a fixed experimental design $P \in \mathcal{E}^{PE}$. We note that each of these four averages is identified, by an experimental arm that occurs with positive probability.

- $\mathbb{E}[Y(0, 0)]$ is identified by the experimental arm $F = (0, 0),\ A^S = (0, 0)$. This arm occurs with positive probability:

  $$P(F = (0, 0), A^S = (0, 0)) = \mathbb{P}(A^S = (0, 0)|F = (0, 0))P(F = (0, 0)) > 0$$

  By, Assumption 1, we know the average outcome in this arm identifies $\mathbb{E}[Y(0, 0)]$:

  $$\mathbb{E}[Y|F = (0, 0), A^S = (0, 0)] = \mathbb{E}[Y(0, 0)|F = (0, 0), A^S = (0, 0)] = \mathbb{E}[Y(0, 0)]$$

  where the second equality comes from our definition of an experimental design (Definition 1), which assumes that the experimental instruments are assigned independently of each individual's latent potential outcomes or potential treatments. In this way, $\mathbb{E}[Y(0, 0)]$ is identified.

- $\mathbb{E}[Y(A_1(1), 0)]$ is identified by two experimental arms: $F = (1, 0)$ and $A^S = (a_1, 0)$ for $a_1 \in \{0, 1\}$. The probability that one of these arms occurs is positive:

  $$P(F = (1, 0), A_2^S = 0) = P(A_2^S = 0|F = (1, 0))P(F = (1, 0)) > 0$$

  And the average outcome in these arms corresponds to the desired average potential outcome (Given Assumption 1 and the randomization guaranteed by the experiment):

  $$\mathbb{E}[Y|F = (1, 0), A_2^S] = \mathbb{E}[Y(A_1(1), 0)|F = (1, 0), A_2^S] = \mathbb{E}[Y(A_1(1), 0)]$$

Thus, $\mathbb{E}[Y(A_1(1), 0)]$ is also identified.

- $\mathbb{E}[Y(0, A_2(0))]$ is identified by two experimental arms: $F = (0, 1)$ and $A^s = (0, a_2)$ for $a_2 \in \{0, 1\}$. The probability that one of these arms occurs is positive:

$$P(F = (0, 1), A_1^s = 0) = P(A_1^s = 0 | F = (0, 1))P(F = (0, 1)) > 0$$

  And the average outcome in these arms corresponds to the desired average potential outcome (Given Assumption 1 and the randomization guaranteed by the experiment):

$$\mathbb{E}[Y|F = (0, 1), A_1^s] = \mathbb{E}[Y(0, A_2(0, 1))|F = (0, 1), A_1^s] = \mathbb{E}[Y(0, A_2(0, 1))]$$

  Thus, $\mathbb{E}[Y(0, A_2(0, 1))]$ is also identified.

- $\mathbb{E}[Y(A_1(1), A_2(A_1(1), 1))]$ is identified by four experimental arms: $F = (1, 1)$ and $A^s = (a_1, a_2)$ for $a_1, a_2 \in \{0, 1\}$. The probability that one of these arms occurs is positive: $P(F = (1, 1)) > 0$. Moreover, the average outcome in these arms corresponds to the desired average potential outcome (Given Assumption 1 and the randomization guaranteed by the experiment):

$$\mathbb{E}[Y|F = (1, 1)] = \mathbb{E}[Y(A_1(1), A_2(A_1(1), 1))|F = (1, 1)] = \mathbb{E}[Y(A_1(1), A_2(A_1(1), 1))]$$

  Thus, $\mathbb{E}[Y(A_1(1), A_2(A_1(1), 1))]$ is also identified.

Thus, $\tilde{V}_t(\lambda_1)$ is identified and in turn its extrema, the attribution bounds are identified.

Step 3. Necessity (If the attribution bounds are identified, then $P \in \mathcal{E}^{PE}$)

Suppose the attribution bounds are identified. We show by contradiction that this requires the design P to be a partial eligibility experiment, i.e., that it must satisfy three conditions: $P(F = f) > 0$ for all $f \in \mathbb{F}$, $P(A^s = (0, 0)|F = (0, 0)) > 0$, and $P(A_t^s = 0|F_t = 0, F_{-t} = 1) > 0$ for $t = 1, 2$. We consider these three conditions in turn.

Step 3.i. Suppose there is an $f \in \mathbb{F}$ such that $P(F = f) = 0$.

(a) Suppose $P(F = (1, 1)) = 0$. Note, that we never observe the joint distribution of $\big(A_1(1),\ A_2(A_1(1), 1),\ Y(A_1(1), A_2(A_1(1), 1))\big)$, and we cannot identify it. To see this, note that we never observe the distribution of $\big(A_1(1), A_2(A_1(1), 1)\big)$, at best, we observe the marginals of $A_1(1)$, $A_2(0, 1)$ and $A_2(1, 1)$, but there are many joint distribution

that respect these marginals. So, for instance, if $A_2(0,1) = 0$ deterministically, but you consider the following two DGPs, you cannot distinguish them, yet they will in general lead to different bounds:

$$\mathbb{P}_B(A_1(1) = b_1, A_2(1,1) = b_2) = 1/4$$
$$\mathbb{P}_C(A_1(1) = b_1, A_2(1,1) = b_2) = \begin{cases} 1/4 + \gamma & \text{if } b_1 = b_2 \\ 1/4 - \gamma & \text{else} \end{cases}$$

Assume, $\gamma \in (0, 1/4)$. In both cases, each random variable has a marginal distribution that sets it equal to 0 or 1 with probability 1/2. However, if $Y(x_1, x_2) = x_1 x_2$, then the bounds in each world boil down to the extrema of $\lambda_1 \mathbb{E}[Y(A_1(1), A_2(A_1(1), 1))] = \lambda_1 \mathbb{E}[A_1(1) A_2(1,1)]]$, which is other $\lambda_1/4$ under $\mathbb{P}_B$ or $\lambda_1/4 + \gamma$ under $\mathbb{P}_C$, and thus the bounds are not identified showing the desired contradiction.

(b) Suppose $P(F = (1,0)) = 0$. Note that we never observe the joint distribution of $A_1(1)$ and $Y(1,0)$, so once again we can construct the same type of counterexample, but now varying the joint distribution over these two variables while holding their marginals fixed. Here, setting the second potential treatments to 1 and only having all the irrelevant potential outcomes set to zero, ensures the bounds just depend on $\mathbb{E}[Y(A_1(1), 0)] = \mathbb{E}[Y(1,0) A_1(1)]$, which under an analogous counterexample to the last one (3.i.a) takes on different values under two observationally equivalent distributions, and is thus not identified. As a consequence, neither are the bounds, proving our desired result.

(c) Suppose $P(F = (0,1)) = 0$. Note that we never observe the joint distribution of $A_2(0,1)$ and $Y(0,1)$, so once again we can construct the same type of counterexample, but now varying the joint distribution over these two variables while holding their marginals fixed. Here, setting $A_1(1) = 1$ and $A_2(1,1) = 1$, while having all the irrelevant potential outcomes set to zero, ensures the bounds just depend on $\mathbb{E}[Y(0, A_2(0,1))] = \mathbb{E}[Y(0,1) A_2(0,1)]$, which under an analogous counterexample to first case (3.i.a) takes on different values under two observationally equivalent distributions, and is thus not identified. As a consequence, neither are the bounds, proving our desired result.

(d) Suppose $P(F = (0,0)) = 0$. Note that if $A_1(1) = 1$ and $A_2(\cdot, 1) = 1$, then with probability one every individual receives at least one treatment, so we never observe the distribution $Y(0,0)$. In particular, fixing the potential outcomes other than $Y(0,0)$ to 0

ensures that the bounds only depend on $\mathbb{E}[Y(0,0)]$, but this can take on any value while being observationally indistinguishable, and so the bounds are unidentified.

Thus, we must have $P(F = f) > 0$ for all $f \in \mathbb{F}$.

Step 3.ii. Suppose that $P(A^s = (0,0)|F = (0,0)) = 0$. Note that if $A_1(1) = 1$ and $A_2(\cdot, 1) = 1$, then with probability one every individual receives at least one treatment, so we never observe the distribution of $Y(0,0)$. As in the last subcase, we can fix the potential outcomes other than $Y(0,0)$ to 0, ensuring that the bounds only depend on $\mathbb{E}[Y(0,0)]$, but this can take on any value while being observationally indistinguishable, and so the bounds are not point-identified. This delivers the desired contradiction.

Step 3.iii. Suppose there is a $t \in \{1,2\}$ such that $P(A_t^s = 0|F_t = 0, F_{-t} = 1) = 0$.

(a) Suppose, $P(A_1^s = 0|F = (0,1)) = 0$. Note that if $A_1(1) = 1$, then we never observe the distribution of $Y(0, A_2(0,1))$, and more to the point, we never observe the distribution of $A_2(0,1)$. To construct two observationally equivalent distributions, suppose the potential outcomes are all 0, unless one receives only the second treatment, in which case it could be 0 or 1, and set $A_2(1,1) = 1$. Here, each term in $\tilde{V}_t(\lambda_1)$ is 0, except for $\mathbb{E}[Y(0, A_2(0,1))] = \mathbb{E}[Y(0,1)A_2(0,1)]$. Even if $Y(0,1) = 1$ deterministically, we cannot identify the average $A_2(0,1)$, so this could be anything in $[0,1]$, and thus the bounds are not point-identified. This delivers the desired contradiction.

(b) Suppose, $P(A_2^s = 0|F = (1,0)) = 0$. Note that if $A_2(\cdot, 1) = 1$, then we never observe the distribution of $Y(A_1(1), 0)$. Suppose that the potential outcomes are always 0, unless only the first treatment is received, and we construct two observationally equivalent distributions, by creating two joint distributions over the first potential treatment, $A_1(1)$ and the relevant potential outcome, $Y(1,0)$, that have the same marginals. In particular, copying the construction from (3.i.a) as applied to (3.i.b) gives us the desired contradiction: we can construct two observationally equivalent distributions that lead to different bounds, and thus, they are not identified.

□

B.4. Proof for Proposition 4

PROPOSITION 4 (continuing from p. 25). *Let $P \in \mathcal{E}^{PE}$ be any partial-eligibility design. There exists an eligibility design $P' \in \mathcal{E}^{E}$ that is weakly more efficient for estimating the four*

single-world counterfactuals. Formally, assuming finite variance for all potential outcomes, the asymptotic variance of any regular and asymptotically linear (RAL) estimator for the means is weakly lower under P' than under P. Moreover, if the true bounds are non-trivial ($\underline{V}_t < \overline{V}_t$), this efficiency gain extends to the estimation of the attribution bounds.

Proof of Proposition 4. The proof proceeds by establishing that the simple sample mean estimator for each single-world counterfactual mean is semiparametrically efficient in this experimental context. The result then follows by comparing the variance of these efficient estimators under the different designs.

Step 1:. Let θ_k for $k \in \{1, 2, 3, 4\}$, denote the expected potential outcome for each of the four single-world counterfactuals. Let Z_{ik} be a binary indicator that is 1 if individual i is assigned to the k-th identifying experimental cell, and 0 otherwise. For example, for $\theta_1 = \mathbb{E}[Y(0,0)]$, the cell is $(F = (0,0), A^s = (0,0))$. The observed outcome for individuals with $Z_{i1} = 1$ is $Y_i = Y_i(0,0)$. The probability of assignment to this cell, $\pi_k = P(Z_{ik} = 1) = p_k$, is known by the design P. By design, the estimand θ_k is the expected outcome under assignment to cell k.

Step 2:. It is a standard result in semiparametric statistics that under standard regularity conditions, the Augmented Inverse Propensity Weighting (AIPW) estimator is semiparametrically efficient for the average potential outcome. In a setting with no covariates, the AIPW estimator for θ_k is given by:

$$\hat{\theta}_k^{\text{AIPW}} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{Z_{ik}}{p_k} (Y_i - \mathbb{E}[Y|Z_{ik} = 1]) + \mathbb{E}[Y|Z_{ik} = 1] \right)$$

In practice, the conditional expectation $\mathbb{E}[Y|Z_{ik} = 1]$ is replaced by a consistent estimator. In this covariate-free setting, the natural and efficient estimator for $\mathbb{E}[Y|Z_{ik} = 1]$ is the sample mean of Y within cell k, which we denote $\hat{\theta}_k = (\sum_i Z_{ik} Y_i)/(\sum_i Z_{ik})$. Substituting this into the AIPW formula:

$$\begin{aligned} \hat{\theta}_k^{\text{AIPW}} &= \frac{1}{N} \sum_{i=1}^{N} \left(\frac{Z_{ik}}{p_k} (Y_i - \hat{\theta}_k) + \hat{\theta}_k \right) \\ &= \frac{1}{N p_k} \sum_{i=1}^{N} Z_{ik} Y_i - \frac{\hat{\theta}_k}{N p_k} \sum_{i=1}^{N} Z_{ik} + \frac{1}{N} \sum_{i=1}^{N} \hat{\theta}_k \end{aligned}$$

$$= \frac{N_k}{N p_k}\hat{\theta}_k - \frac{N_k}{N p_k}\hat{\theta}_k + \hat{\theta}_k = \hat{\theta}_k$$

where $N_k = \sum_i Z_{ik}$. This shows that the AIPW estimator is algebraically identical to the simple sample mean within the cell. Since the AIPW estimator is known to be semiparametrically efficient, it follows that the simple sample mean, $\hat{\theta}_k$, is an efficient estimator for θ_k in this experimental context.

Step 3: . The asymptotic variance of an efficient estimator is given by the semiparametric efficiency bound. For estimating a mean from a subpopulation sampled with probability p_k, this bound is σ_k^2/p_k, where $\sigma_k^2 = \text{Var}(Y_i|Z_{ik})$. The estimators for each θ_k are based on disjoint samples and are thus independent. The asymptotic variance matrix for the vector of efficient estimators $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_4)^T$ under design P is therefore:

$$\Sigma_P = \frac{1}{N}\text{diag}\left(\frac{\sigma_1^2}{p_1}, \frac{\sigma_2^2}{p_2}, \frac{\sigma_3^2}{p_3}, \frac{\sigma_4^2}{p_4}\right)$$

Let P' be the corresponding eligibility design constructed from P. The variance matrix under design P' is:

$$\Sigma_{P'} = \frac{1}{N}\text{diag}\left(\frac{\sigma_1^2}{p_1'}, \frac{\sigma_2^2}{p_2'}, \frac{\sigma_3^2}{p_3'}, \frac{\sigma_4^2}{p_4'}\right)$$

As established previously, P' reallocates probability from non-informative exogenous arms, ensuring $p_k' \geq p_k$ for $k = 1, 2, 3, 4$. Thus, every diagonal element of Σ_P is weakly greater than the corresponding element of $\Sigma_{P'}$. This means the variance of the efficient estimator for each θ_k is weakly lower under the eligibility design P'.

Step 4: Extension to Attribution Bounds.. The extension to the attribution bounds $(\underline{V}_t, \overline{V}_t)$ follows from the Delta Method. When the true bounds are non-trivial ($\underline{V}_t < \overline{V}_t$), the bounds are smooth (continuously differentiable) functions of the vector of means θ. The asymptotic variance of an efficient estimator for a smooth function of parameters is a quadratic form involving the variance matrix of the efficient estimators for those parameters. Since the variance matrix $\Sigma_{P'}$ is smaller than $\mathbf{\Sigma}_P$ (i.e., $\mathbf{\Sigma}_P - \mathbf{\Sigma}_{P'}$ is positive semidefinite), the resulting asymptotic variance for the efficient estimator of the bounds will also be weakly smaller under design P'. This completes the proof. □

REMARK A1. *Two clarifications are in order. First, the optimality statement rules out improvement by regular estimators that use only cell membership and outcomes; it does not*

rule out variance reduction achievable by using pre-treatment covariates. Second, irregular/superefficient estimators can outperform at a subset of parameter space but are not uniformly better, and are excluded by the regularity assumption above.

B.5. Proof for Lemma 2

THEOREM 2 (continuing from p. 27). *The estimators, $\hat{\underline{V}}_t$ and $\hat{\overline{V}}_t$, are consistent. Furthermore, under standard regularity conditions for the central limit theorem, $\sqrt{N}(\Lambda_t\hat{\mu} - \Lambda_t\mu)$ converges to a multivariate normal distribution, $\mathcal{N}\left(0, \Lambda_t\Sigma\Lambda_t^T\right)$, as N grows, where $\Sigma = Cov(Z_i)$ and $\mu = \mathbb{E}[Z_i]$. The bounds themselves, $\hat{\underline{V}}_t$ and $\hat{\overline{V}}_t$, converges either to the min/max of two normals, $\Lambda_t\hat{\mu}$, when the bounds collapse (i.e., $\underline{V}_t = \overline{V}_t$), or to a normal centered around the corresponding bound when the bounds do not collapse (i.e., $\underline{V}_t < \overline{V}_t$).*

Proof of Theorem 2. The proof is presented in two parts: first, we establish consistency of the estimators, and second, we characterize their asymptotic distribution.

Part 1: Consistency. Consistency is established using the Law of Large Numbers (LLN) and the Continuous Mapping Theorem (CMT).

First, we show that the Horvitz-Thompson estimator for each single-world mean, $\hat{\mu}_f$, is consistent for the true mean, which we denote μ_f. The estimator is $\hat{\mu}_f = N^{-1}\sum_{i=1}^{N} Z_{i,f}$, where the IPW outcome $Z_{i,f}$ is defined in Estimator 1. The terms $Z_{i,f}$ are i.i.d. across i. We show that $\mathbb{E}[Z_{i,f}] = \mu_f$. Let $p_f = P(F_i = f, A_i^s \cdot (1 - F_i) = 0)$ be the known probability of assignment to the identifying cell for μ_f.

$$
\begin{aligned}
\mathbb{E}[Z_{i,f}] &= \mathbb{E}\left[\frac{\mathbf{1}\{F_i = f, A_i^s \cdot (1 - F_i) = 0\}}{p_f} Y_i\right] \\
&= \mathbb{E}\left[Y_i | F_i = f, A_i^s \cdot (1 - F_i) = 0\right] \\
&= \begin{cases}
\mathbb{E}[Y(0,0)] & \text{if } f = (0,0) \\
\mathbb{E}[Y(A_1(1),0)] & \text{if } f = (1,0) \\
\mathbb{E}[Y(0,A_2(0,1))] & \text{if } f = (0,1) \\
\mathbb{E}[Y(A_1(1),A_2(A_1(1),1))] & \text{else}
\end{cases}
\end{aligned}
$$

Since $Z_{i,f}$ are i.i.d. and under standard regularity conditions have finite expectation μ_f, the Law of Large Numbers implies that $\hat{\mu}_f \xrightarrow{p} \mu_f$. This holds for all $f \in \mathbb{F}$. Therefore, the vector of estimators $\hat{\mu}$ is consistent for the vector of true means μ: $\hat{\mu} \xrightarrow{p} \mu$.

The candidate values for the bounds are $\Lambda_t\hat{\mu}$. Since matrix multiplication is a continuous function, the CMT implies $\Lambda_t\hat{\mu} \xrightarrow{p} \Lambda_t\mu$. Let $(\tilde{V}_t(0), \tilde{V}_t(1)) = \Lambda_t\mu$ be the true candidate values.

Finally, the estimators for the bounds are $\underline{\hat{V}}_t = \min(\Lambda_t\hat{\mu})$ and $\overline{\hat{V}}_t = \max(\Lambda_t\hat{\mu})$. The min / max functions are continuous, so a further application of the CMT yields:

$$\underline{\hat{V}}_t \xrightarrow{p} \min(\Lambda_t\mu) = \underline{V}_t$$
$$\overline{\hat{V}}_t \xrightarrow{p} \max(\Lambda_t\mu) = \overline{V}_t$$

Thus, the estimators for the attribution bounds are also consistent.

Part 2: Asymptotic Distribution. The characterization of the asymptotic distribution relies on the Multivariate Central Limit Theorem (CLT) and the Delta Method, generalized for non-differentiable functions.

First, we apply the Multivariate CLT to the vector of estimated means $\hat{\mu}$. As $\hat{\mu}$ is the sample average of the i.i.d. random vectors $Z_i = (Z_{i,00}, \ldots, Z_{i,11})^T$, and assuming finite second moments, the CLT implies:

$$\sqrt{N}(\hat{\mu} - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

where $\mu = \mathbb{E}[Z_i]$ and $\Sigma = \mathrm{Cov}(Z_i) = \mathbb{E}[Z_i Z_i^T] - \mu\mu^T$. The (f, f') element of Σ is $\mathrm{Cov}(Z_{i,f}, Z_{i,f'})$. Note that for $f \neq f'$, the product $Z_{i,f}Z_{i,f'}$ is identically zero because an individual cannot be in two different identifying cells simultaneously. Thus, for $f \neq f'$, $\mathbb{E}[Z_{i,f}Z_{i,f'}] = 0$ and $\mathrm{Cov}(Z_{i,f}, Z_{i,f'}) = -\mu_f\mu_{f'}$.

Next, consider the linear transformation $\Lambda_t\hat{\mu}$. By the properties of multivariate normal distributions, we have:

$$\sqrt{N}(\Lambda_t\hat{\mu} - \Lambda_t\mu) = \Lambda_t\left(\sqrt{N}(\hat{\mu} - \mu)\right) \xrightarrow{d} \mathcal{N}(0, \Lambda_t\Sigma\Lambda_t^T)$$

Let $(G_0, G_1)^T \sim \mathcal{N}(0, \Lambda_t\Sigma\Lambda_t^T)$ denote the limiting random vector.

Finally, we characterize the distribution of the bounds estimators. Let's analyze the lower bound $\underline{\hat{V}}_t$.

$$\sqrt{N}(\underline{\hat{V}}_t - \underline{V}_t) = \sqrt{N}\left(\min(\hat{V}_t(0), \hat{V}_t(1)) - \min(V_t(0), V_t(1))\right)$$

We consider two cases, as stated in the theorem.

- *Case 1: The bounds are distinct ($\underline{V}_t < \overline{V}_t$).* This implies $V_t(0) \neq V_t(1)$. Without loss of generality, assume $V_t(0) < V_t(1)$, so $\underline{V}_t = V_t(0)$. Since $\hat{V}_t(0)$ and $\hat{V}_t(1)$ are consistent, for large N, we will have $\hat{V}_t(0) < \hat{V}_t(1)$ with probability approaching one. In this event, $\hat{\underline{V}}_t = \hat{V}_t(0)$. Therefore, the asymptotic distribution of $\sqrt{N}(\hat{\underline{V}}_t - \underline{V}_t)$ is identical to that of $\sqrt{N}(\hat{V}_t(0) - V_t(0))$. This corresponds to the first component of the limiting normal vector, G_0. Thus, $\sqrt{N}(\hat{\underline{V}}_t - \underline{V}_t) \xrightarrow{d} G_0 \sim \mathcal{N}(0, (\Lambda_t \Sigma \Lambda_t^T)_{11})$. The distribution is normal. The same argument applies for the joint distribution of the bounds.

- *Case 2: The bounds collapse ($\underline{V}_t = \overline{V}_t$).* This implies $V_t(0) = V_t(1) = \underline{V}_t$. In this case, we can no longer assume that one estimator will be smaller than the other. We have:

$$\begin{aligned}\sqrt{N}(\hat{\underline{V}}_t - \underline{V}_t) &= \sqrt{N}(\min(\hat{V}_t(0), \hat{V}_t(1)) - V_t(0)) \\ &= \min\left(\sqrt{N}(\hat{V}_t(0) - V_t(0)), \sqrt{N}(\hat{V}_t(1) - V_t(1))\right)\end{aligned}$$

  By the Continuous Mapping Theorem, since the vector $(\sqrt{N}(\hat{V}_t(0) - V_t(0)), \sqrt{N}(\hat{V}_t(1) - V_t(1)))^T$ converges in distribution to $(G_0, G_1)^T$, the expression above converges in distribution to $\min(G_0, G_1)$. This is the minimum of two (potentially correlated) normal random variables, which is a non-normal distribution. The same argument applies for the joint distribution of the bounds.

This completes the proof. □

B.6. Proof for Algorithm 1

THEOREM A1 (Validity of Subsampling for Attribution Bounds). *Suppose the conditions of Theorem 2 hold. For $t \in \{1, 2\}$, let $\hat{\theta}_{N,t}$ be either the lower bound estimator $\hat{\underline{V}}_t$ or the upper bound estimator $\hat{\overline{V}}_t$, calculated on the full sample of size N. Let θ_t be its corresponding true value. Let b be a subsample size such that $b \to \infty$ and $b/N \to 0$ as $N \to \infty$. Then, the subsampling-based confidence interval for θ is asymptotically valid.*

PROOF. The proof proceeds by verifying the conditions of the general subsampling theorem of Politis and Romano (1994). The main condition is that the normalized statistic, $\sqrt{N}(\hat{\theta}_{N,t} - \theta_t)$, must converge in distribution to some limit J. However, we know from Theorem 2 that this converges either to a normal distribution or to the max of two normals, depending on whether or not the bounds collapse (i.e., $\underline{V}_t = \overline{V}_t$).

In both cases, $\sqrt{N}(\hat{\theta}_{N,t} - \overline{V}_t)$ converges to a well-defined limiting distribution J. The conditions of the subsampling theorem are met. Therefore, the quantiles of the

subsampling distribution for $\sqrt{b}(\hat{\theta}^*_{t,b} - \hat{\theta}_{N,t})$ provide a consistent estimate of the quantiles of the true sampling distribution of $\sqrt{N}(\hat{\theta}_{N,t} - \theta_t)$. This allows for the construction of an asymptotically valid confidence interval. The logic applies for the upper and lower bounds as well as for each $t \in \{1, 2\}$. □

B.7. Proof for Proposition 5

PROPOSITION 5 (continuing from p. 29). *Under Assumptions 1, 3 and 4, all six counterfactuals means required for the attribution bounds (Lemma 2) are identified and can be expressed in terms of observable distributions:*

$$\begin{aligned}
\mathbb{E}[Y(0,0)] &= \mathbb{E}\left[\mathbb{E}[Y|A=(0,0),X]\right] \\
\mathbb{E}[Y(A_1,0)] &= \sum_{a_1} \mathbb{E}\left[\mathbb{E}[Y|A=(a_1,0),X]\mathbb{P}(A_1=a_1|X)\right] \\
\mathbb{E}[Y(0,A_2(0))] &= \sum_{a_2} \mathbb{E}\left[\mathbb{E}[Y|A=(0,a_2),X]\mathbb{P}(A_2=a_2|A_1=0,X)\right] \\
\mathbb{E}[Y(0,A_2(A_1))] &= \sum_{a_2} \mathbb{E}\left[\mathbb{E}[Y|A=(0,a_2),X]\mathbb{P}(A_2=a_2|X)\right] \\
\mathbb{E}[Y(A_1,A_2(0))] &= \sum_{a_1,a_2} \mathbb{E}\left[\mathbb{E}[Y|A=(a_1,a_2),X]\mathbb{P}(A_1=a_1|X)\mathbb{P}(A_2=a_2|A_1=0,X)\right] \\
\mathbb{E}[Y(A_1,A_2)] &= \mathbb{E}[Y]
\end{aligned}$$

As a consequence, the attribution bounds, $\underline{V}_t$ and $\overline{V}_t$, are also identified for each t.

Proof of Proposition 5. Throughout we assume Assumptions 1 (the Causal Model), 3 (Sequential Unconfoundedness), and 4 (Positivity) hold. We begin by deriving a few general identification results, which help simplify our subsequent proof that the six required quantities are identified. For clarity in the proof, we use the simplified observational notation $A_1 = A_1(1)$ and $A_2(a_1) = A_2(a_1, 1)$.

General Identification Result.. Here, we start by showing that two useful conditional expectations are identified.

First, we show that the conditional expectation $\mathbb{E}[Y(a_1, a_2)|X = x]$ is identified for

any fixed treatment path (a_1, a_2).

(eq. 1)
$$\begin{aligned}
&\mathbb{E}[Y(a_1, a_2)|X = x] \\
&\quad = \mathbb{E}[Y(a_1, a_2)|A_1 = a_1, X = x] && \text{Ass. 3} \\
&\quad = \mathbb{E}[Y(a_1, a_2)|A_1 = a_1, A_2(a_1) = a_2, X = x] && \text{Ass. 3} \\
&\quad = \mathbb{E}[Y|A_1 = a_1, A_2 = a_2, X = x] && \text{Ass. 1}
\end{aligned}$$

This result is the standard g-formula for a fixed treatment path, which identifies the unobservable potential outcome mean with an observable conditional mean. Here, assumption 4, ensures that the conditioning event in the final expression $\{A_1 = a_1, A_2 = a_2, X = x\}$, occurs with non-zero probability for all relevant values of (a_1, a_2, x). Without positivity, the conditional expectation $\mathbb{E}[Y|A_1 = a_1, A_2 = a_2, X = x]$ would not be identified. All subsequent identification results rely on Assumption 4 in an analogous manner.

Second, we identify the conditional expectation, $\mathbb{E}[Y(a_1, a_2)|A_2(\tilde{a}_1) = a_2, X = x]$ for any $a_1, a_2, \tilde{a}_1 \in \{0, 1\}$:

(eq. 2)
$$\begin{aligned}
&\mathbb{E}[Y(a_1, a_2)|A_2(\tilde{a}_1) = a_2, X = x] \\
&\quad = \mathbb{E}[Y(a_1, a_2)|A_1 = \tilde{a}_1, A_2(\tilde{a}_1) = a_2, X = x] && \text{Ass. 3} \\
&\quad = \mathbb{E}[Y(a_1, a_2)|A_1 = \tilde{a}_1, X = x] && \text{Ass. 3} \\
&\quad = \mathbb{E}[Y(a_1, a_2)|A_1 = a_1, X = x] && \text{Ass. 3} \\
&\quad = \mathbb{E}[Y(a_1, a_2)|A_1 = a_1, A_2(a_1) = a_2, X = x] && \text{Ass. 3} \\
&\quad = \mathbb{E}[Y|A_1 = a_1, A_2 = a_2, X = x] && \text{Ass. 1}
\end{aligned}$$

Identification of the Six Counterfactual Means..

(a) $\mathbb{E}[Y(0, 0)]$: Here, we apply the Law of Iterated Expectations (abbreviated, L.I.E.) over the distribution of X, and then apply eq. 1:

$$\begin{aligned}
\mathbb{E}[Y(0, 0)] &= \mathbb{E}\big[\mathbb{E}[Y(0, 0)|X]\big] && \text{L.I.E.} \\
&= \mathbb{E}\big[\mathbb{E}[Y|A_1 = 0, A_2 = 0, X]\big] && \text{eq. 1}
\end{aligned}$$

(b) $\mathbb{E}[Y(A_1, 0)]$: Using the Law of Iterated Expectations over the distribution of A_1 and X for the first two equalities, and applying Assumption 3 for the third equality and

finally plugging in our earlier results gives us the following:

$$\begin{aligned}
\mathbb{E}[Y(A_1, 0)] &= \mathbb{E}\Big[\mathbb{E}[Y(A_1, 0)|X]\Big] && \text{L.I.E.} \\
&= \mathbb{E}\left[\sum_{a_1} \mathbb{E}[Y(a_1, 0)|A_1 = a_1, X]P(A_1 = a_1|X)\right] && \text{L.I.E.} \\
&= \mathbb{E}\left[\sum_{a_1} \mathbb{E}[Y(a_1, 0)|X]P(A_1 = a_1|X)\right] && \text{Ass. 3} \\
&= \mathbb{E}\left[\sum_{a_1} \mathbb{E}[Y|A_1 = a_1, A_2 = 0, X]P(A_1 = a_1|X)\right] && \text{eq. 1}
\end{aligned}$$

(c) $\mathbb{E}[Y(0, A_2(0))]$: Once again, use the Law of Iterated Expectations twice (over X and $A_2(0)$) to get the first equality, apply Assumption 3 for the second equality, and then use Assumption 1 to get the desired result:

$$\begin{aligned}
&\mathbb{E}[Y(0, A_2(0))] \\
&= \mathbb{E}\left[\sum_{a_2} \mathbb{E}[Y(0, a_2)|A_2(0) = a_2, X]P(A_2(0) = a_2|X)\right] && \text{L.I.E.} \\
&= \mathbb{E}\left[\sum_{a_2} \mathbb{E}[Y(0, a_2)|A_1 = 0, A_2(0) = a_2, X]P(A_2(0) = a_2|A_1 = 0, X)\right] && \text{Ass. 3} \\
&= \mathbb{E}\left[\sum_{a_2} \mathbb{E}[Y|A_1 = 0, A_2 = a_2, X]P(A_2 = a_2|A_1 = 0, X)\right] && \text{Ass. 1}
\end{aligned}$$

(d) $\mathbb{E}[Y(0, A_2(A_1))]$: Once again, use the Law of Iterated Expectations twice (over X and $A_2(A_1)$) to get the first equality, apply Assumption 3 for the third equality, and use Assumption 1 to get the second and final equalities:

$$\begin{aligned}
&\mathbb{E}[Y(0, A_2(A_1))] \\
&= \mathbb{E}\left[\sum_{a_2} \mathbb{E}[Y(0, a_2)|A_2(A_1) = a_2, X]\mathbb{P}(A_2(A_1) = a_2|X)\right] && \text{L.I.E.} \\
&= \mathbb{E}\left[\sum_{a_2} \mathbb{E}[Y(0, a_2)|A_2 = a_2, X]\mathbb{P}(A_2 = a_2|X)\right] && \text{Ass. 1} \\
&= \mathbb{E}\left[\sum_{a_2} \mathbb{E}[Y(0, a_2)|A_1 = 0, A_2 = a_2, X]\mathbb{P}(A_2 = a_2|X)\right] && \text{Ass. 3}
\end{aligned}$$

$$= \mathbb{E}\left[\sum_{a_2} \mathbb{E}[Y|A_1 = 0, A_2 = a_2, X]P(A_2 = a_2|X)\right] \qquad \text{Ass. 1}$$

(e) $\mathbb{E}[Y(A_1, A_2(0))]$: Once again, use the Law of Iterated Expectations twice (over X and A_1), then apply Assumption 3

$$\begin{aligned}
&\mathbb{E}[Y(A_1, A_2(0))] \\
&= \mathbb{E}\left[\sum_{a_1} \mathbb{E}[Y(a_1, A_2(0))|A_1 = a_1, X]\mathbb{P}(A_1 = a_1|X)\right] && \text{L.I.E.} \\
&= \mathbb{E}\left[\sum_{a_1} \mathbb{E}[Y(a_1, A_2(0))|X]\mathbb{P}(A_1 = a_1|X)\right] && \text{Ass. 3} \\
&= \mathbb{E}\left[\sum_{a_1,a_2} \mathbb{E}[Y(a_1, a_2)|A_2(0) = a_2, X]\mathbb{P}(A_2(0) = a_2|X)\mathbb{P}(A_1 = a_1|X)\right] && \text{L.I.E.} \\
&= \mathbb{E}\left[\sum_{a_1,a_2} \mathbb{E}[Y|A_1 = a_1, A_2 = a_2, X]\mathbb{P}(A_2 = a_2|A_1 = 0, X)\mathbb{P}(A_1 = a_1|X)\right] && \text{eq. 2 with } \tilde{a}_1 = 0
\end{aligned}$$

(f) $\mathbb{E}[Y(A_1, A_2)]$: By Assumption 1, $Y(A_1, A_2(A_1)) = Y$. So, this is the expected observed outcome:

$$\mathbb{E}[Y(A_1, A_2(A_1))] = \mathbb{E}[Y] \qquad \text{Ass. 1}$$

All six required counterfactual means are expressed in terms of observable conditional expectations and probabilities, and are therefore identified under the stated assumptions. □

B.8. Proof for Theorem A2

THEOREM A2. *Under Assumptions 1, 3 and 4 and standard regularity conditions for estimating parametric models, the vector of six estimated counterfactual means is consistent and asymptotically normal. Consequently, the vector of possible attribution values, $\hat{V}_t = (\hat{V}_t(\lambda))_{\lambda\in\{0,1\}^3}$, is also consistent and asymptotically normal. The estimator for the lower (upper) bound, $\underline{\hat{V}}_t$, is asymptotically normal if the minimum (maximum) of $V_t(\lambda)$*

is unique.[13] *Otherwise, its asymptotic distribution is that of the minimum (maximum) of multiple normal random variables.*

Proof of Theorem A2. The proof establishes the consistency and asymptotic normality of the vector of the six estimated counterfactual means, which we denote $\hat{\theta}$. The results for the final attribution bounds then follow via the Continuous Mapping Theorem and the Delta Method.

The proof relies on standard M-estimation theory for two-step estimators. The use of K-fold cross-fitting is a crucial feature that simplifies the asymptotic analysis, by ensuring that the estimation error from the first-stage nuisance models does not affect the first-order asymptotic distribution of the final estimator, provided the nuisance models are correctly specified.

Part 1: Consistency. We first establish the consistency of the estimator vector $\hat{\theta}$ for the true vector of counterfactual means θ_0.

(a) Under standard regularity conditions for parametric M-estimation (e.g., for logistic or linear regression), the estimator for the nuisance parameters, $\hat{\gamma}^{(-k)}$, computed on data outside of fold k, is a consistent estimator for the true parameter vector γ_0. That is, $\hat{\gamma}^{(-k)} \xrightarrow{p} \gamma_0$ as the sample size $N(1 - 1/K) \to \infty$.

(b) For each observation i in a given fold k, the predicted values for the nuisance functions (e.g., $\hat{\mu}_i(a_1, a_2) = \hat{\mu}^{(-k)}(a_1, a_2, X_i; \hat{\gamma}^{(-k)})$) are constructed using an estimator $\hat{\gamma}^{(-k)}$ that is independent of the observation's own data (Y_i, A_i, X_i).

(c) Since the parametric functions $\mu(\cdot)$ and $p(\cdot)$ are continuous in the parameters γ, and $\hat{\gamma}^{(-k)}$ is consistent, the cross-fitted predictions are consistent estimators of the true conditional functions. For instance, $\hat{\mu}_i(a_1, a_2)$ converges in probability to the true function $\mu(a_1, a_2, X_i)$ for all $a_1, a_2 \in \{0, 1\}$. The same applies for the propensities, $\hat{p}_i(a_2|a_1)$, which converges in probability to the true conditional probability, $p(a_2|a_1, X_i)$, and $\hat{p}_i(a_1)$, which converges in probability to the true conditional probability $p(a_1|X_i)$ (for all $a_1, a_2 \in \{0, 1\}$).

(d) Each component of $\hat{\theta}$ is a sample average of these predicted values or a sample average of the product of predicted values. As an example of the first case, consider

[13] Letting $V_{t,(1)} \leq V_{t,(2)} \leq \ldots \leq V_{t,(8)}$ denote the order statistics for the vector of true values $\{V_t(\lambda)\}_{\lambda \in \{0,1\}^3}$, the condition for a unique minimum is $V_{t,(1)} < V_{t,(2)}$, and for a unique maximum is $V_{t,(8)} > V_{t,(7)}$.

the first component, $\hat{\theta}_1 = N^{-1} \sum_{i=1}^{N} \hat{\mu}_i(0, 0)$. By a uniform law of large numbers, this sample average converges in probability to the expectation of the true function:

$$\hat{\theta}_1 = \frac{1}{N} \sum_{i=1}^{N} \hat{\mu}_i(0, 0) \xrightarrow{p} \mathbb{E}[\mu(0, 0, X_i)] = \mathbb{E}[\mathbb{E}[Y|A = (0, 0), X]] = \theta_{1,0}$$

As an example of the second case, consider the second component:

$$\hat{\theta}_2 = \sum_{a_1} N^{-1} \sum_{i=1}^{N} \hat{\mu}_i(a_1, 0)\, \hat{p}_i(a_1)$$

To show that this converges in probability to $\theta_2 = \sum_{a_1} \mathbb{E}[\mu(a_1, 0, X)\, p(a_1|X)]$, we break down the error for a fixed a_1 into an approximation and a sampling error:

$$\begin{aligned}
&\frac{1}{N} \sum_i \hat{\mu}_i(a_1, 0)\, \hat{p}_i(a_1) - \mathbb{E}[\mu(a_1, 0, X)\, p(a_1|X)] \\
&\quad = \underbrace{\left(\frac{1}{N} \sum_i \hat{\mu}_i(a_1, 0)\, \hat{p}_i(a_1) - \mu_i(a_1, 0)\, p_i(a_1) \right)}_{\text{Approximation Error}} \\
&\qquad + \underbrace{\left(\frac{1}{N} \sum_i \mu_i(a_1, 0)\, p_i(a_1) - \mathbb{E}[\mu(a_1, 0, X)\, p(a_1|X)] \right)}_{\text{Sampling Error}}
\end{aligned}$$

The sampling error vanishes because of the law of large numbers, while the first term goes to zero because of the individual estimates are consistent and Slutsky's theorem ensures their product is as well. In this way, we see that the error for a fixed a_1 goes to zero, and thus, our estimator is consistent. This logic applies to all six components of the estimator. Therefore, the vector estimator $\hat{\theta}$ is consistent for the true vector θ_0.

Part 2: Asymptotic Normality. To establish asymptotic normality, we derive the influence function for $\hat{\theta}$. A key result from the literature on two-step estimation with cross-fitting (e.g., Chernozhukov et al., 2018, "Double/debiased machine learning for treatment and structural parameters") is that for correctly specified nuisance models, the first-stage estimation error is asymptotically negligible. The estimator behaves as if

the true nuisance parameters γ_0 were known.

The estimator $\hat{\theta}$ is asymptotically equivalent to the "oracle" estimator that uses the true nuisance functions:

$$\sqrt{N}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\boldsymbol{m}(X_i; \gamma_0) - \theta_0) + o_p(1)$$

where $\boldsymbol{m}(X_i; \gamma_0)$ is the vector of identification formulas from Proposition 5 evaluated at the true parameters. For example, the first component is $m_1(X_i; \gamma_0) = \mu(0, 0, X_i; \gamma_0)$.

Let the influence function for $\hat{\theta}$ be the vector $\psi(W_i) = \boldsymbol{m}(X_i; \gamma_0) - \theta_0$, where $W_i = (Y_i, A_i, X_i)$. The components of this influence function are:

$$\begin{aligned}
\psi_1(W_i) &= \mu(0, 0, X_i) - \theta_{1,0} \\
\psi_2(W_i) &= \sum_{a_1} \mu(a_1, 0, X_i)\, p(a_1 | X_i) - \theta_{2,0} \\
&\vdots \\
\psi_6(W_i) &= Y_i - \theta_{6,0}
\end{aligned}$$

The terms $\psi(W_i)$ are i.i.d. random vectors with $\mathbb{E}[\psi(W_i)] = \mathbb{E}[\boldsymbol{m}(X_i; \gamma_0)] - \theta_0 = \theta_0 - \theta_0 = \mathbf{0}$. Assuming finite second moments of these terms, the Multivariate Central Limit Theorem applies:

$$\sqrt{N}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(W_i) + o_p(1) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Omega})$$

where the asymptotic variance-covariance matrix is $\boldsymbol{\Omega} = \text{Var}(\psi(W_i)) = \mathbb{E}[\psi(W_i)\psi(W_i)^T]$. This establishes the asymptotic normality of the vector of the six estimated counterfactual means.

Part 3: Asymptotic Distribution of the Bounds.. The vector of possible attribution values, $\hat{\boldsymbol{V}}_t$, is a linear transformation of $\hat{\theta}$. Let $\hat{\boldsymbol{V}}_t = \boldsymbol{L}_t \hat{\theta}$ for the appropriate 8×6 transformation matrix $\boldsymbol{L}_t$ that maps the six means to the eight candidate bound values (one for each $\lambda \in \{0, 1\}^3$). By the Delta Method, $\sqrt{N}(\hat{\boldsymbol{V}}_t - \boldsymbol{V}_t)$ is also asymptotically normal with mean zero and covariance matrix $\boldsymbol{L}_t \boldsymbol{\Omega} \boldsymbol{L}_t^T$.

The estimator for the lower bound is $\underline{\hat{V}}_t = \min(\hat{\boldsymbol{V}}_t)$. Its asymptotic distribution depends on the uniqueness of the minimum of the true vector $\boldsymbol{V}_t$.

- *Case 1: Unique Minimum.* If the minimum element of $\boldsymbol{V}_t$ is unique (i.e., $V_{t,(1)} < V_{t,(2)}$

where $V_{t,(s)}$ denotes the s order statistic for the vector V_t), then with probability approaching one, $\hat{\underline{V}}_t$ will be equal to the single component of the vector $\hat{\boldsymbol{V}}_t$ that corresponds to this minimum. In this case, the asymptotic distribution of $\sqrt{N}(\hat{\underline{V}}_t - \underline{V}_t)$ is simply the marginal normal distribution of that component.

- *Case 2: Non-Unique Minimum.* If multiple components of $\boldsymbol{V}_t$ share the same minimum value, then the asymptotic distribution of $\sqrt{N}(\hat{\underline{V}}_t - \underline{V}_t)$ is given by the minimum of the corresponding components of the limiting multivariate normal random vector. This distribution is non-normal.

A symmetric argument holds for the upper bound estimator $\hat{\overline{V}}_t = \max(\hat{\boldsymbol{V}}_t)$. This completes the proof. □