# Cause and Effect in Antidiscrimination Law

*Hillel J. Bavli*[*]

*ABSTRACT: Standards of causation in antidiscrimination law, and disparate-treatment cases in particular, are deeply flawed. Their defects have caused an illogical, obscure, and unworkable proof scheme that requires an overhaul to curb the harm that it engenders and to allow the antidiscrimination statutes to serve their objectives effectively. This Article proposes a theory and method of causation that achieves this goal. The problem stems from the inadequacies associated with current standards of causation in disparate-treatment cases—the but-for test and the motivating-factor test. The proposed "factorial" approach introduces a causal standard that addresses these inadequacies. It entails three innovations over current causation schemes: (1) it adopts a predominant framework for cause and effect in the sciences, called the "potential-outcomes" framework, as a central structure in which to sharply define and analyze the causal inquiry; (2) it employs a causal measure, called the "NESS" test, that refines and, in a sense, unifies the but-for and motivating-factor tests by retaining the central feature of the but-for test—the "necessity condition"—but in a less restrictive form; and (3) it applies a legal framework grounded in tort law and recent advances regarding multiple sufficient causes. In addition to reflecting actual cause and effect, the proposed approach promotes antidiscrimination law's deterrence and fairness objectives, and it allows an interpretation of causal language in antidiscrimination statutes that is consistent with good policy and Congress's intent—an interpretation not possible under current standards.*

[*] Assistant Professor of Law, SMU Dedman School of Law. The author wishes to thank Barry Goldstein, Ellen Katz, Nina Varsava, Tirthankar Dasgupta, Daniel Hemel, Edward Cheng, Grant Hayden, Donald Rubin, and the participants of the Michigan Law School 2020 Junior Scholars Conference for their helpful comments and analysis, and Sophie Franks and Laura Sundin for their valuable editing and research assistance.

## I. INTRODUCTION

Standards of causation in antidiscrimination law, and disparate-treatment cases in particular, are deeply flawed. They can be described, in the words of Dr. Seuss, as a "muddled duddled fuddled wuddled fox in socks."[1] That's a metaphor for causation doctrine that entails not only a tangled web of inconsistent rules and conditions but also causal standards that are not quite right in the first instance. The problem stems from the inadequacies associated with current standards of causation in disparate-treatment cases —the "but-for" test and the "motivating-factor" test.

Causation is the element of a legal claim that connects a defendant's misconduct to a plaintiff's injury.[2] As the Supreme Court has stated, "When the law grants persons the right to compensation for injury from wrongful conduct, there must be some demonstrated connection, some link, between the injury sustained and the wrong alleged. The requisite relation between prohibited conduct and compensable injury is governed by the principles of causation . . . ."[3]

The most basic and pervasive standard of causation in the law is the but-for standard.[4] It asks whether the outcome would have occurred in the absence of the alleged conduct.[5] The defendant's conduct is a but-for cause of the plaintiff's harm if, in the absence of the alleged conduct, the plaintiff's harm would not have occurred.[6] The but-for test is generally simple and provides a straightforward method for determining actual cause and effect.[7] It is based on the "necessity condition"—on whether the defendant's conduct "made a difference."

In disparate-treatment cases, however, the but-for test is frequently inadequate as a standard of causation. This is because most disparate-treatment cases involve allegations of at least two possible factors (or "mixed motives") leading to an adverse employment action—one illegitimate (i.e., discriminatory) and one legitimate.[8] This occurs for two reasons. First, employment decisions, such as whom to hire, fire, or promote, or how much to pay an employee, are complex, often involving multiple factors. Second, it

[1] DR. SEUSS, FOX IN SOCKS 59 (1993).

[2] Comcast Corp. v. Nat'l Ass'n of Afr. Am.-Owned Media, 140 S. Ct. 1009, 1013 (2020); Univ. of Tex. Sw. Med. Ctr. v. Nassar, 570 U.S. 338, 342 (2013).

[3] *Nassar*, 570 U.S. at 342.

[4] DAN B. DOBBS, PAUL T. HAYDEN & ELLEN M. BUBLICK, HORNBOOK ON TORTS 317 (2d ed. 2016).

[5] *Id.*

[6] *Id.*

[7] Throughout this Article, I use the terms "cause and effect" or "causal" effect to refer to the generic causal concept, "causation" to refer to the legal causal concept, and "causality" or "causal inference" to refer to the scientific or statistical causal concept.

[8] *See* BARBARA T. LINDEMANN, PAUL GROSSMAN & C. GEOFFREY WEIRICH, EMPLOYMENT DISCRIMINATION LAW 2-2–3 (5th ed. 2012).

is difficult to ascertain an employer's true motive, and, therefore, an employer can easily rebut an allegation of discrimination by presenting evidence of a legitimate purpose.

Disparate-treatment cases, and mixed-motive cases in particular, can often fail to satisfy the but-for standard of causation since, even in the absence of the discriminatory factor, the legitimate factor would be sufficient for the adverse employment action. Consider, for example, circumstances in which an employer promoted a male employee over a female employee named Lisa, leading to allegations of unlawful discrimination based on sex. Assume there is strong evidence that (1) the employer held the general view that male employees are more capable than female employees and looked specifically to promote a male employee, but (2) Lisa was, incidentally, far less qualified for the position than was her male colleague. Employing the but-for standard would result in a finding of no causation and therefore no liability because even in the absence of the discriminatory conduct, the employer would not have hired Lisa for the position.

Courts and legislatures have struggled immensely with this result for at least two reasons. First, based on common intuition regarding cause and effect, it is near-universally recognized that, if a factor is sufficient and necessary to bring about an outcome on its own, then it is a cause of the outcome even if other factors would themselves be sufficient to bring about the same outcome. Second, allowing discriminatory employers to avoid liability by identifying a legitimate factor (whether pretextual or not) is arguably contrary to the deterrence objectives of antidiscrimination law. Moreover, the question of motive is rarely straightforward evidentiarily, and the problem of defining causation in mixed-motive cases is closely tied to difficulties in ascertaining an employer's true purpose and the ease with which an employer can contrive a legitimate explanation for an adverse employment action.

As a result of the inadequacy of the but-for test in disparate-treatment cases, courts and legislatures have frequently sought to replace the but-for test with the motivating-factor test, which simply requires that discrimination with respect to a protected characteristic somehow motivated an adverse employment action, whether the discriminatory factor actually made a difference or not.[9] This test is, however, inadequate as a replacement for the but-for standard. It does not (and is not intended to) reflect actual cause and effect. Its meaning is vague, and it simply allows a factfinder to rely on intuition to determine whether an employer should be held liable for the adverse employment decision based on evidence of discrimination, regardless of whether a causal link has been established.

[9] *Id.* at 2-112–14.

Courts and legislatures therefore choose between two inadequate causation standards. Consequently, disparate-treatment doctrine, frequently analogized to "a swamp,"[10] has become increasingly chaotic—rife with complexities, inconsistencies, and ambiguities—as courts have applied varied approaches to counteract these inadequacies in order to arrive at desirable case outcomes. For example, courts apply one standard to actions under Title VII, for discrimination based on "race, color, religion, sex, or national origin,"[11] a different standard to actions under the Age Discrimination in Employment Act (ADEA), for discrimination based on age,[12] and under Title VII's retaliation provision, for retaliation against an employee who challenged discrimination,[13] and a mix of standards to actions under the Americans with Disabilities Act (ADA), for discrimination based on disability.[14] Moreover, the various standards entail a complex web of intricacies that reflect the inadequacies of the causation measures that courts apply.

In this Article, I introduce a causal framework that returns logic and clarity to the disparate-treatment analysis. This framework, which I call the "factorial" framework,[15] provides a straightforward method for determining causation in disparate-treatment cases, and one that fulfills the policy objectives of antidiscrimination law. As importantly, it provides a theory of causation that is firmly grounded in the actual, common, and scientific notion of cause and effect, and that permits a more seamless relationship between *standards* of causation and *proof* of causation.

Employing the factorial framework as an underlying theory of causation leads to three important innovations over current approaches to determining causation in disparate-treatment cases. First, the framework imports a counterfactual model of cause and effect and its conceptual building blocks from the "potential-outcomes" model (also known as the "Rubin Causal Model"), a predominant model for asking and answering questions of cause and effect in statistics and the sciences. In particular, the factorial framework places the but-for test within this broader counterfactual model, the broader reasoning on which it is based. This is crucial for understanding and applying causal measures, or "estimands," that are broader than the but-for measure but retain its fundamental feature—the necessity condition. As such, it is also fundamental to a more appropriate causal estimand in disparate-treatment

[10] Martin J. Katz, *Unifying Disparate Treatment (Really)*, 59 HASTINGS L.J. 643, 645 n.8 (2008) ("Courts and commentators have routinely referred to current disparate treatment doctrine as a 'swamp,' a 'morass,' and a 'quagmire.'" (citing Costa v. Desert Palace, Inc., 299 F.3d 838, 851–53 (9th Cir. 2002), *aff'd*, 539 U.S. 90 (2003), and other sources)).

[11] 42 U.S.C. § 2000e-2 (2018).

[12] 29 U.S.C. § 623 (2018).

[13] 42 U.S.C. § 2000e-3.

[14] 42 U.S.C. § 12112.

[15] *See* Hillel J. Bavli, *Counterfactual Causation*, 51 ARIZ. ST. L.J. 879, 904 (2019) [hereinafter *Counterfactual Causation*].

cases. Moreover, the potential-outcomes model provides building blocks of causal inference that refine the causal question and simplify concepts that have caused substantial confusion in these cases. It also facilitates a causal inquiry that permits a stronger link between the law's measure of causation and proof thereof. The potential-outcomes model enables a factfinder to better conceptualize the causal problem, and, specifically, the inferences required for determining causation. This is especially useful in discrimination cases, in which a complex set of factors may underlie an adverse employment decision.

Second, the factorial framework applies the NESS (Necessary Element of a Sufficient Set) test as the broader causal estimand appropriate in disparate-treatment cases. This test is based on the notion of a "causal set," which can be understood as a set of factors that together lead to the occurrence of an outcome.[16] It is satisfied if a factor "was necessary for the sufficiency of a set of existing antecedent conditions that was sufficient for the occurrence of the consequence."[17] A particular and simplified version of this test has been adopted by the Restatement (Third) of Torts as a test for causation in situations involving multiple sufficient causes (msc). It provides that, "[i]f multiple acts occur, each of which under § 26 [Factual Cause] alone would have been a factual cause of the physical harm at the same time in the absence of the other act(s), each act is regarded as a factual cause of the harm."[18]

As I will argue, in addition to reflecting the actual, common, and scientific notion of cause and effect, the NESS test simplifies the causal inquiry in disparate-treatment cases and fulfills the policy objectives of antidiscrimination law. In applying NESS as the causal estimand—the measure of interest—in these cases, the factorial framework can be understood as refining and unifying the but-for and motivating-factor tests of causation. Indeed, the factorial framework treats the NESS test as a middle-ground approach that retains the essential element of the but-for test—that, in line with common, scientific, and philosophic notions of cause and effect, the misconduct "made a difference," or, in some sense, satisfied the necessity condition—while capturing the essential purpose of the motivating-factor

[16] RESTATEMENT (THIRD) OF TORTS: LIAB. FOR PHYSICAL AND EMOTIONAL HARM § 26 cmt. c (AM. L. INST. 2010) ("A useful model for understanding factual causation is to conceive of a set made up of each of the necessary conditions for the plaintiff's harm. Absent any one of the elements of the set, the plaintiff's harm would not have occurred."); *id.* § 27 cmt. f (explaining the "multiple-sufficient-causal-set situation"); *see also* June v. Union Carbide Corp., 577 F.3d 1234, 1242–44 (10th Cir. 2009); *infra* Section III.E.

[17] Richard W. Wright, *The Grounds and Extent of Legal Responsibility*, 40 SAN DIEGO L. REV. 1425, 1441 (2003) [hereinafter Wright, *Grounds and Extent*]; Richard W. Wright, *Once More into the Bramble Bush: Duty, Causal Contribution, and the Extent of Legal Responsibility*, 54 VAND. L. REV. 1071, 1102–03 (2001) [hereinafter Wright, *Once More into the Bramble Bush*].

[18] RESTATEMENT (THIRD) OF TORTS: LIAB. FOR PHYSICAL AND EMOTIONAL HARM § 27 (AM. L. INST. 2010).

test—to hold an employer responsible for discrimination that "motivated" an adverse employment action, even if it was accompanied by sufficient legitimate motivating factors. Importantly, however, the factorial framework should not be understood as a compromise approach; rather, I argue that it *refines* existing tests and more accurately captures the notion of actual cause and effect than either test does individually.

Third, the factorial framework can be applied simply and straightforwardly using basic torts principles. I treat mixed-motive disparate-treatment cases as a special type of msc torts situation, a well-studied situation in which two or more forces contribute to an outcome where each force alone would be sufficient to produce the same outcome.[19] Additionally, I apply basic torts principles for establishing a prima facie claim and satisfying burdens of production and persuasion. The refinements allowed by the factorial framework's application of the potential-outcomes model and the NESS test obviate the need for complicated burden-shifting rules and other overly complex features of current proof schemes. Instead, simple torts principles involving proof of a wrong, a harm, and a causal link between the wrong and the harm suffice.

The factorial framework carries a wide range of implications for antidiscrimination law. In this Article, I discuss two sets of implications in particular: implications for the policy objectives of antidiscrimination law and implications for judicial interpretation of causal language in antidiscrimination statutes. I argue that, in addition to reflecting actual cause and effect, the proposed approach promotes antidiscrimination law's deterrence objectives while preventing windfall recoveries and their distorting effects on incentives. It also allows an interpretation of causal language in antidiscrimination statutes that is consistent with good policy and Congress's intent—an interpretation not possible under current standards.

In Part II, I provide an overview of causation doctrine surrounding mixed-motive disparate-treatment cases. I discuss the inconsistent and inadequate standards governing these cases and their relationship to msc situations in tort law. In Part III, I introduce fundamental theory and concepts underlying the factorial framework and their applicability in disparate-treatment cases. Among other things, I introduce the potential-outcomes model and apply it to explain how the factorial framework refines both the but-for test and the motivating-factor test, and why NESS is an ideal causal estimand in disparate-treatment cases. In Part IV, I show how the factorial framework would apply in practice in disparate-treatment cases. I provide illustrations and show how the factorial framework could simplify and

[19] *See id.*; W. PAGE KEETON, DAN B. DOBBS, ROBERT E. KEETON & DAVID G. OWEN, PROSSER AND KEETON ON THE LAW OF TORTS § 41, at 266 (5th ed. 1984) ("[T]here is one type of situation in which [the but-for test] fails. If two causes concur to bring about an event, and either one of them, operating alone, would have been sufficient to cause the identical result, some other test is needed.").

improve the disparate-treatment analysis. In Part V, I discuss implications for the policy aims of antidiscrimination law and for interpreting causal language in antidiscrimination statutes. In Part VI, I conclude.

## II. CAUSATION IN DISPARATE-TREATMENT CASES

### *A. SCRAMBLING FOR A LOGICAL CAUSAL STANDARD*

Title VII of the Civil Rights Act of 1964 prohibits discrimination against an individual "because of such individual's race, color, religion, sex, or national origin."[20] In *Price Waterhouse v. Hopkins*,[21] six Justices agreed—though not in a majority opinion—on a burden-shifting framework in which, first, a plaintiff alleging discrimination would "show that one of the prohibited traits was a 'motivating' or 'substantial' factor in the employer's decision," and, second, if the plaintiff satisfied her burden, "the burden of persuasion would shift to the employer, which could escape liability if it could prove that it would have taken the same employment action in the absence of all discriminatory animus"—i.e., that discrimination was not a but-for cause of the adverse employment action.[22] According to the plurality opinion in *Price Waterhouse*:

> We need not leave our common sense at the doorstep when we interpret a statute. It is difficult for us to imagine that, in the simple words "because of," Congress meant to obligate a plaintiff to identify the precise causal role played by legitimate and illegitimate motivations in the employment decision she challenges. We conclude, instead, that Congress meant to obligate her to prove that the employer relied upon sex-based considerations in coming to its decision.[23]

The opinion, however, emphasized that the "preservation of an employer's remaining freedom of choice"—the "maintenance of employer prerogatives"—"is evident from the statute itself and from its history, both in Congress and in this Court."[24] The Court concluded, therefore, that "an employer shall not be liable if it can prove that, even if it had not taken [the prohibited trait] into account, it would have come to the same decision regarding a particular person."[25]

[20] 42 U.S.C. § 2000e-2(a) (2018).

[21] Price Waterhouse v. Hopkins, 490 U.S. 228 (1989).

[22] Univ. of Tex. Sw. Med. Ctr. v. Nassar, 570 U.S. 338, 348 (2013) (citing *Price Waterhouse*, 490 U.S. at 258 (plurality opinion); *id.* at 259–60 (White, J., concurring); *id.* at 276–77 (O'Connor, J., concurring)).

[23] *Price Waterhouse*, 490 U.S. at 241–42 (plurality opinion).

[24] *Id.* at 242.

[25] *Id.*

The method that emerged from *Price Waterhouse* was based on the Supreme Court's decision in *McDonnell Douglas Corp. v. Green*,[26] a case frequently understood as the starting point of the law's current proof scheme in Title VII cases. In *McDonnell Douglas*, the Court held that a Title VII plaintiff may establish a prima facie discrimination claim

> by showing (i) that he belongs to a racial minority; (ii) that he applied and was qualified for a job for which the employer was seeking applicants; (iii) that, despite his qualifications, he was rejected; and (iv) that, after his rejection, the position remained open and the employer continued to seek applicants from persons of complainant's qualifications.[27]

This test has appeared in various forms since *McDonnell Douglas*; however,

> [r]egardless of the form of the prima facie case . . . if the plaintiff is relying on the *McDonnell Douglas* methodology, he generally must in some way prove four elements: (1) membership in a protected class; (2) qualification for the job; (3) an adverse employment action; and (4) a causal connection between the adverse action and protected classification.[28]

If the plaintiff establishes a prima facie discrimination claim, under *McDonnell Douglas*, "[t]he burden then must shift to the employer to articulate some legitimate, nondiscriminatory reason for the employee's rejection."[29] At that point, however, the plaintiff is given "a fair opportunity to show that [the employer's] stated reason for [the plaintiff's] rejection was in fact pretext," meaning "that the presumptively valid reasons for his rejection were in fact a coverup for a racially discriminatory decision."[30]

Notably, *McDonnell Douglas* and its progeny require only weak circumstantial evidence of causation to establish a prima facie case.[31] Once this standard has been met, the burden shifts to the defendant to prove that discriminatory conduct is not the but-for cause of the alleged adverse employment action.[32] This proof scheme may reflect, on the one hand, the Court's understanding of Title VII as requiring a but-for standard of causation and, on the other hand, the Court's realization that requiring a plaintiff to

[26] McDonnell Douglas Corp. v. Green, 411 U.S. 792 (1973).

[27] *Id.* at 802.

[28] LINDEMANN ET AL., *supra* note 8, at 2-13.

[29] *McDonnell Douglas*, 411 U.S. at 802.

[30] *Id.* at 804–05.

[31] *See id.* at 802.

[32] *Id.*

prove but-for causation to establish a prima facie claim would be impracticable and would defeat the statute's objectives.[33]

Finally, because employment decisions are complex, and it is often feasible for an employer to contrive a legitimate purpose for an adverse employment decision, once the employer has proved such a purpose, the plaintiff is then given the opportunity to show that the legitimate reason provided by the employer is in fact pretextual.[34] Evidence showing that the employer's reason is pretextual can consist of "direct" evidence, such as discriminatory comments by an employer, as well as statistical evidence and evidence showing that similarly-situated employees without the protected feature were treated more favorably than the plaintiff.[35]

In 1991, two years after *Price Waterhouse*, Congress amended Title VII.[36] It "codified the burden-shifting and lessened causation framework of *Price Waterhouse* in part but also rejected it to a substantial degree."[37] The amendment added a subsection providing that "an unlawful employment practice is established when the complaining party demonstrates that race, color, religion, sex, or national origin was a motivating factor for any employment practice, even though other factors also motivated the practice."[38] However, it replaced the Court's burden-shifting framework with a new one that allowed an employer to avoid "damages or . . . an order requiring any admission, reinstatement, hiring, promotion, or payment" if it proved that the illegitimate motivating factor was not a but-for cause of the adverse employment action, but did not allow the employer to avoid "declaratory relief, injunctive relief . . ., and [certain] attorney's fees and costs" with such a showing.[39] Therefore, pursuant to Title VII, courts apply a version of the proof scheme applied in *McDonnell Douglas* and *Price Waterhouse*, but with the remedial framework outlined in its 1991 amendments.

The Supreme Court again addressed "the meaning of 'because' and the problem of causation"[40] in 2009, in *Gross v. FBL Financial Services, Inc.*[41] *Gross* involved the ADEA, which prohibits discrimination against an individual "because of such individual's age."[42] According to the Court, "[t]he words 'because of' mean 'by reason of: on account of'" and, therefore, "the ordinary meaning of the ADEA's requirement that an employer took adverse action

[33] *See infra* Section II.B; *McDonnell Douglas*, 411 U.S. at 800–06.

[34] *McDonnell Douglas*, 411 U.S. at 804–05.

[35] *See* LINDEMANN ET AL., *supra* note 8, at 2-50–102.

[36] Univ. of Tex. Sw. Med. Ctr. v. Nassar, 570 U.S. 338, 348 (2013).

[37] *Id.*

[38] 42 U.S.C. § 2000e-2(m) (2018).

[39] *Id.* § 2000e-5(g)(2)(B); *see Nassar*, 570 U.S. at 348–49.

[40] *Nassar*, 570 U.S. at 349.

[41] Gross v. FBL Fin. Servs., Inc., 557 U.S. 167 (2009).

[42] 29 U.S.C. § 623(a) (2018).

'because of' age is that age was the 'reason' that the employer decided to act."[43] The Court held that "[t]o establish a disparate-treatment claim under the plain language of the ADEA, therefore, a plaintiff must prove that age was the 'but-for' cause of the employer's adverse decision."[44] The Court explicitly rejected the application of Title VII's burden-shifting framework to prove causation under the ADEA, holding "that under § 623(a)(1), the plaintiff retains the burden of persuasion to establish that age was the 'but-for' cause of the employer's adverse action."[45] It reasoned that the ADEA, unlike Title VII, does not provide for the motivating-factor test, highlighting that "Congress neglected to add such a provision to the ADEA when it amended Title VII . . . even though it contemporaneously amended the ADEA in several ways."[46]

In 2013, the Supreme Court, in *University of Texas Southwestern Medical Center v. Nassar*,[47] revisited the issue of causation under Title VII, but this time under Title VII's antiretaliation provision, which prohibits discrimination against an individual "because he has opposed any practice made an unlawful employment practice by this subchapter, or because he has made a charge, testified, assisted, or participated in any manner in an investigation, proceeding, or hearing under this subchapter."[48] The Court again interpreted the term "because" to imply a but-for standard. It held that, "[g]iven the lack of any meaningful textual difference between the text in this statute and the one in *Gross*, the proper conclusion here, as in *Gross*, is that Title VII retaliation claims require proof that the desire to retaliate was the but-for cause of the challenged employment action."[49]

Finally, as in Title VII and the ADEA, the ADA uses language that gives rise to complex issues of causation in mixed-motive cases. The statute prohibits discrimination against "individual[s] on the basis of disability in regard to job application procedures, the hiring, advancement, or discharge of employees, employee compensation, job training, and other terms, conditions, and privileges of employment."[50] As with other antidiscrimination statutes, there is little stability or consistency when it comes to the standard of causation in mixed-motive ADA claims. For example, some circuits apply the but-for test, thus precluding liability when an illegitimate factor is accompanied by sufficient legitimate factors, while others apply the

[43] *Gross*, 557 U.S. at 176 (quoting 1 WEBSTER'S THIRD NEW INTERNATIONAL DICTIONARY 194 (1966)).

[44] *Id.*

[45] *Id.* at 177.

[46] *Id.* at 174.

[47] Univ. of Tex. Sw. Med. Ctr. v. Nassar, 570 U.S. 338 (2013).

[48] 42 U.S.C. § 2000e-3(a) (2018).

[49] *Nassar*, 570 U.S. at 352; *see Counterfactual Causation*, *supra* note 15, at 927–30.

[50] 42 U.S.C. § 12112(a).

motivating-factor test, which allows a plaintiff to establish a claim even when the discriminatory factor is not a but-for cause of the adverse decision.[51]

In summary, causal standards in antidiscrimination law consist of a complex maze of rules and conditions rife with inconsistencies, both within and across discrimination statutes, and within and across jurisdictions. Moreover, as I will argue, the tortuous routes through this maze lead only to vague and unsuitable causal measures. Indeed, the "swamp" that describes the current state of the law surrounding causation in these cases has likely grown from the inadequacy of these measures.

### *B. The Underlying Problem: Mixed-Motive Cases as Multiple-Sufficient-Cause Situations*

Inconsistency in standards surrounding disparate-treatment cases is only a symptom of a larger underlying problem. The source of the inconsistency is the inadequacy of the but-for and motivating-factor tests of causation.

A mixed-motive case is a type of multiple-sufficient-cause (msc) situation. This situation arises when at least two forces contribute to an outcome and each alone would have produced the same outcome. For example, consider a hypothetical case involving two fires, each negligently and independently started, that combine to burn down a lodge, where each fire alone would be sufficient to destroy the lodge—even in the absence of the other fire.[52] These fires are called concurrent multiple sufficient causes because they occurred concurrently with respect to the harm.[53] A variant of this problem is one in which one fire arrives at the lodge immediately after the first fire burned the lodge down. The fires in this circumstance are called *successive* multiple sufficient causes because they occurred successively with respect to the harm.[54] Concurrent and successive msc situations are also known as "duplicative" and "preemptive" causation problems.[55] In this Section, I focus on concurrent msc situations; I turn to the issue of successive msc situations in later sections of the Article.

A mixed-motive case is a special type of msc situation because, like the general case, it involves at least two factors, at least one legitimate factor and one discriminatory factor, that contribute to an outcome—an adverse

[51] *See* LINDEMANN ET AL., *supra* note 8, at 13-197–99 (discussing variation in standards across circuits and over time, and citing cases).

[52] *See generally* Kingston v. Chi. & N.W. Ry. Co., 211 N.W. 913 (Wis. 1927) (involving "two separate, independent, and distinct [fires], each of which constituted the proximate cause of plaintiff's damage, and either of which, in the absence of the other, would have accomplished such result"); DOBBS ET AL., *supra* note 4, at 321.

[53] *See generally* WARD FARNSWORTH & MARK F. GRADY, TORTS: CASES AND QUESTIONS 293 (3d ed. 2019) (discussing msc situations).

[54] *See generally id.* (using the term "subsequent").

[55] *See* Richard W. Wright, *Causation in Tort Law*, 73 CALIF. L. REV. 1735, 1775–77 (1985).

employment action—where each alone would be sufficient to produce the same outcome.[56] Consider, for example, a case involving a racist employer who rejected the employment application of an African-American applicant, the plaintiff, based on the employer's preference for white workers. At the same time that the plaintiff was applying for the position, a white candidate also applied. Incidentally, this white applicant had been training for this position for years and was better qualified for it. Regardless of who else applied for the position—white or not—it is clear that the employer would likely have hired this applicant over the plaintiff. There are two forces that prevented the employer from hiring the African-American applicant: the employer's discriminatory treatment and the qualifications of the African-American applicant relative to those of the white applicant. Each force was sufficient on its own to prevent the plaintiff's hire.

As in the general msc problem, neither factor in the mixed-motive case is a but-for cause of the harm under the traditional but-for test because even in the absence of the employer's discriminatory hiring process, the African-American applicant would not have been hired. The discrimination, in a certain sense, made no difference at all with respect to the outcome of the hiring process since the employer would have hired the highly qualified candidate over other candidates, regardless of their race.

To be sure, mixed-motive cases are arguably unique in an important respect relative to the general class of msc situations: they often involve circumstances in which at least one of the two or more forces sufficient for the adverse employment action is due to either "nature" or non-protected features or behavior of the plaintiff, rather than a third-party source of misconduct.[57] For example, in the illustration above, the application of the highly qualified white candidate was a sufficient factor, but one arising from nature rather than another individual's misconduct. Other examples of factors that are due to either nature or a plaintiff's own non-protected features or behavior are: an employer's decision to shut down a business; the plaintiff's own poor qualifications for a position; the plaintiff's poor performance or rudeness to customers; and a downturn in the economy causing an employer to reduce her staff or to hire fewer employees.

[56] *See Counterfactual Causation*, *supra* note 15, at 882–83, 927–32 (treating mixed-motive discrimination cases as a particular type of multiple-sufficient-cause situation); Andrew Verstein, *The Failure of Mixed-Motives Jurisprudence*, 86 U. CHI. L. REV. 725, 742–54, 755 (2019) (challenging "rationales sometimes given for incorporating causation into a mixed-motives standard" and arguing that, even "[i]f we analogize a defendant's motives as potential causes of the action, then mixed motives are analogous to torts with multiple causes").

[57] Some scholars have questioned whether motives should be understood as causes or assessed with respect to causation in the first instance. *See* Verstein, *supra* note 56, at 742–46. This Article accepts as a premise Congress's and the Supreme Court's treatment of motives as forces that qualify to be causes of an outcome and that are subject to principles of causation. *See supra* Section II.A.

However, this feature of mixed-motive cases is not necessarily distinct from the general msc situation and, in any event, should not have a substantial impact on our analysis, except to highlight that an appropriate standard of causation must account for concerns regarding behavioral effects broader than simply deterring discriminatory conduct. In particular, it must account for behavioral effects associated with inappropriate findings of liability and windfall recoveries.

Msc situations, like mixed-motive cases in particular, involve at least one factor arising from misconduct (e.g., a negligently started fire or a discriminatory purpose), but involve other sufficient factors that can arise from either third-party misconduct (e.g., a second negligently started fire), natural forces (a second fire arising from a lightning strike and dry conditions), or the plaintiff's own (non-protected) features or misconduct (e.g., a second fire arising from the plaintiff's negligence). While mixed-motive cases could arise from two separate sources of discriminatory conduct, they usually arise within one of the latter two categories in particular—meaning, they ordinarily involve sufficient forces that arise from nature or the plaintiff's own non-protected features or behavior. This is arguably significant because courts deciding torts cases have sometimes treated the latter two categories of msc situations differently than the first category. Some courts—such as the Supreme Court of Wisconsin in the famous *Kingston* case—have held that, although a plaintiff can recover in a msc situation even though the defendant's misconduct is not a but-for cause of the harm, a defense to this general rule is that the other sufficient force was due to nature rather than another's misconduct.[58] Similarly, where a msc situation involves the plaintiff's own misconduct as a sufficient factor, the plaintiff's recovery may be foreclosed by rules of contributory or comparative negligence.[59]

However, although the classic two-fire problem arises from the misconduct of two independent parties, many msc situations outside of the mixed-motive context involve "natural" or "contributory" factors. It is true that a drag racer who sues another drag racer for negligence, where each drag racer's reckless behavior is deemed to be sufficient for the occurrence of the harm, is far different than an applicant who sues for discrimination even though his qualifications for the position were substantially worse than those expected for the position. One can argue that it is, in a sense, the applicant's "fault" that he is not a better applicant; but even if the applicant's qualifications can be described as his "fault," this meaning of fault is certainly different than in the drag racing context.

[58] Kingston v. Chi. & N.W. Ry. Co., 211 N.W. 913, 914–15 (Wis. 1927) ("From our present consideration of the subject, we are not disposed to criticise [sic] the doctrine which exempts from liability a wrongdoer who sets a fire which unites with a fire originating from natural causes, such as lighting, not attributable to any human agency, resulting in damage.").

[59] *See generally* DOBBS ET AL., *supra* note 4, at 379–408 (discussing plaintiff misconduct).

Nevertheless, other msc situations in the torts context are more analogous to this mixed-motive scenario. Consider, for example, a case in which it is alleged that a defendant tortiously exposed a plaintiff to a toxin that caused the plaintiff to suffer from a certain illness, but where other factors, such as genetics or other exposures that were not necessarily the "fault" of the plaintiff, may also have been sufficient to bring about the same illness. Similarly, in other mixed-motive cases, a non-discriminatory factor may be more analogous to a contributory-negligence scenario—for example, in cases in which an employee is fired for drinking on the job, regularly arriving late to work, or assaulting a customer.

Suffice it to say, a non-discriminatory factor in a mixed-motive case can arise from natural sources, third-party sources, or from the plaintiff's own non-protected features or behavior. This is true of the general msc situation as well. Even if the mixed-motive context can be said to have a higher proportion of cases that involve natural sources or the plaintiff's own non-protected features or behavior, the mixed-motive context invokes similar concerns as the general msc context, except perhaps with special consideration for socially undesirable incentives that can arise from windfall recoveries and inappropriate findings of liability. Let us, therefore, consider the causal standards that courts have applied in msc situations as they pertain to the mixed-motive context.

Msc situations have given rise to widespread confusion and controversy surrounding the appropriate standard of causation in a broad range of areas.[60] In the concurrent msc situation, neither factor is a but-for cause of the harm. In the two-fire problem, for example, neither fire is a but-for cause of the lodge's destruction since, absent either fire, the same damage would have occurred as a result of the other fire.[61] As a leading treatise states, "The but-for test in such cases leads to a result that is almost always condemned as violating both an intuitive sense of causation and good legal policy."[62] While the traditional but-for test leads to the conclusion that neither fire in the two-fire problem is a cause of the lodge's destruction, "[o]ur senses have told us that [the defendant] *did* participate. . . . In the language of the layman, the defendant's fire 'had something to do with' the burning of plaintiff's property."[63] As one author has commented, in msc situations, "the but-for test denies the existence of cause in fact while everything in human experience and intuition cries out that cause in fact was present."[64]

[60] *See Counterfactual Causation*, *supra* note 15, at 884–93.

[61] *See* DOBBS ET AL., *supra* note 4, at 321.

[62] *Id.* at 321–22.

[63] Wex S. Malone, *Ruminations on Cause-in-Fact*, 9 STAN. L. REV. 60, 89 (1956).

[64] David W. Robertson, *The Common Sense of Cause in Fact*, 75 TEX. L. REV. 1765, 1777 (1997).

Courts and scholars have therefore almost universally declared that the counterfactual model of causation, and the but-for test in particular, fails in msc situations.[65] Courts have abandoned the but-for test in msc situations, and the asserted failure of the counterfactual model in these situations has led to calls to abandon it as a general model of causation.[66]

In msc situations, courts have generally followed the Restatement (Second) of Torts and replaced the but-for test with the "substantial-factor" test, which is a general version of the motivating-factor test.[67] The substantial-factor test asks simply whether the misconduct at issue was "a substantial factor in bringing about the [plaintiff's] harm."[68] The standard does not invoke (and is not intended to invoke) any analytical reasoning derived from models of actual cause and effect; rather, it relies on the intuition of the factfinder to arrive at a determination regarding factual causation. For this reason, it has been said that "[t]he substantial factor test is not so much a test as an incantation."[69] In particular, the substantial-factor test, like the motivating-factor test,

> points neither to any reasoning nor to any facts that will assist courts or lawyers in resolving the question of causation. Put differently, the substantial factor test requires no particular mental operation. It invites the jury's intuition. In one view, that represents a loss of precision in analysis with no corresponding gain.[70]

In contrast with the substantial-factor test, the but-for test of causation reflects notions of cause and effect in philosophy and the sciences, as well as the meaning of cause and effect as commonly used.[71] It also provides a

[65] *See* DOBBS ET AL., *supra* note 4, at 321–22; KEETON ET AL., *supra* note 19, § 41, at 266.

[66] *Counterfactual Causation*, *supra* note 15, at 882; *see, e.g.*, Mitchell v. Gonzales, 819 P.2d 872, 876 (Cal. 1991) (rejecting the but-for test in favor of the substantial-factor test in msc situations); MICHAEL S. MOORE, CAUSATION AND RESPONSIBILITY: AN ESSAY IN LAW, MORALS, AND METAPHYSICS 411 (2009).

[67] RESTATEMENT (SECOND) OF TORTS § 431 (AM. L. INST. 1965); *see* Burrage v. United States, 571 U.S. 204, 215–16 (2014) ("One prominent authority on tort law asserts that 'a broader rule . . . has found general acceptance: The defendant's conduct is a cause of the event if it was a material element and a substantial factor in bringing it about.'" (quoting KEETON ET AL., *supra* note 19, § 41, at 267)).

[68] RESTATEMENT (SECOND) OF TORTS § 431 (AM. L. INST. 1965).

[69] DOBBS ET AL., *supra* note 4, at 323.

[70] *Id.*

[71] *See Burrage*, 571 U.S. at 210–14; Paroline v. United States, 572 U.S. 434, 452 (2014) (referring to alternatives to the but-for test as "a kind of legal fiction or construct"); Gross v. FBL Fin. Servs., Inc., 557 U.S. 167, 176–78 (2009) (holding that "the ordinary meaning of the ADEA's requirement that an employer took adverse action 'because of' age is that the age was the 'reason' that the employer decided to act," and that this implies that, "under the plain language of the ADEA . . . a plaintiff must prove that age was the 'but-for' cause of the employer's adverse decision"); *see also Counterfactual Causation*, *supra* note 15, at 880 n.1, 884–85; DOBBS ET AL., *supra* note 4, at 314.

straightforward analytical process for the factfinder to employ to arrive at a determination regarding causation. But, as the discussion above indicates, it leads—somewhat paradoxically—to incorrect results in msc situations: "despite the seemingly sound and well-accepted reasoning of the counterfactual model, and the but-for standard in particular, applying this reasoning leads to a seemingly illogical conclusion."[72] Relatedly, the but-for test leads to results in msc situations that are commonly recognized as contrary to the deterrence and fairness objectives of tort law.[73]

In the mixed-motive context, the but-for test and the motivating-factor test give rise to the same concerns as in the general msc situation. The but-for test yields results that are counter to our common intuition regarding cause and effect and that are contrary to the deterrence and fairness objectives of antidiscrimination law.[74] For example, if an employer fires an employee because he is African American, intuition regarding cause and effect, as well as good policy, would prevent the employer from defending his conduct on the grounds that he would have fired the employee anyway.[75]

As in other msc contexts, therefore, courts and lawmakers have sought to replace the but-for test in disparate-treatment cases with an alternative test. They have turned to the motivating-factor test, a variation of the substantial-factor test adapted for the antidiscrimination context.[76]

However, the motivating-factor test is subject to the same criticisms as the substantial-factor test. It does not reflect actual cause and effect.[77] It relies on

[72] *Counterfactual Causation*, *supra* note 15, at 885.

[73] DOBBS ET AL., *supra* note 4, at 321–22.

[74] *See infra* Section V.A.

[75] *See infra* Section V.A.

[76] *See, e.g.*, Mt. Healthy City Sch. Dist. Bd. of Educ. v. Doyle, 429 U.S. 274, 287 (1977) ("Initially, in this case, the burden was properly placed upon respondent to show that his conduct was constitutionally protected, and that this conduct was a 'substantial factor' or to put it in other words, that it was a 'motivating factor' in the Board's decision not to rehire him." (footnote omitted)); *see also* Price Waterhouse v. Hopkins, 490 U.S. 228, 259 (1989) (White, J., concurring) ("And here, as in *Mt. Healthy*, and as the Court now holds, Hopkins was not required to prove that the illegitimate factor was the only, principal, or true reason for petitioner's action. Rather, . . . her burden was to show that the unlawful motive was a *substantial* factor in the adverse employment action."); *id.* at 280 (Kennedy, J., dissenting) ("The shift in the burden of persuasion occurs only where a plaintiff proves by direct evidence that an unlawful motive was a substantial factor actually relied upon in making the decision.").

[77] *See* Univ. of Tex. Sw. Med. Ctr. v. Nassar, 570 U.S. 338, 342 (2013) ("When the law grants persons the right to compensation for injury from wrongful conduct, there must be some demonstrated connection, some link, between the injury sustained and the wrong alleged. The requisite relation between prohibited conduct and compensable injury is governed by the principles of causation . . . ."); James E. Viator, *When Cause-in-Fact is More than a Fact: The Malone-Green Debate on the Role of Policy in Determining Factual Causation in Tort Law*, 44 LA. L. REV. 1519, 1526–27 (1984) ("A common belief . . . is that policy considerations have no role to play in the determination of cause-in-fact, 'because no policy can be strong enough to warrant the imposition of liability for loss to which the defendant's conduct has not *in fact* contributed.'"

intuition rather than providing a factfinder with particular direction or an analytical process by which to arrive at a causal determination.[78] And, because it relies on intuition, it produces unpredictability in an area of the law in which such unpredictability may have significant negative consequences on employer and employee behavior, in addition to increasing litigation costs.[79]

For these reasons, courts and lawmakers have been reluctant to apply the motivating-factor test in mixed-motive cases. A central idea in tort law, of which antidiscrimination law is a special type,[80] is that defendants should be held responsible to pay only for damage that they caused.[81] But courts and lawmakers understand that employing the motivating-factor test means allowing juries to impose liability and damages based on intuition regarding who should be held responsible for the harm that has occurred, without a finding of causation—at least in the common or scientific meaning of the term.[82] Moreover, for reasons described in detail below, there is a justifiable fear of overdeterrence and other harmful effects that could result from allowing jurors to decide liability based on a loose and poorly defined standard.[83]

Courts and lawmakers have thus struggled to decide between two inadequate standards of causation in disparate-treatment cases—as they have in other msc situations. They have sought to retain a rigorous standard that reflects actual cause and effect while, at the same time, departing from such a standard, the but-for standard, for purposes of achieving desired outcomes in

(quoting JOHN G. FLEMING, THE LAW OF TORTS 170 (6th ed. 1983))). The tradition of requiring causation, as currently defined, is largely grounded in theories of fairness. But, while some scholars have commented that causation is not necessary for optimal deterrence, few deny that causation, in at least *some* meaningful sense, is necessary for proper levels of deterrence. *See generally* WILLIAM M. LANDES & RICHARD A. POSNER, THE ECONOMIC STRUCTURE OF TORT LAW 229 (1987); William M. Landes & Richard A. Posner, *Causation in Tort Law: An Economic Approach*, 12 J. LEGAL STUD. 109 (1983); Steven Shavell, *An Analysis of Causation and the Scope of Liability in the Law of Torts*, 9 J. LEGAL STUD. 463 (1980); Guido Calabresi, *Concerning Cause and the Law of Torts: An Essay for Harry Kalven, Jr.*, 43 U. CHI. L. REV. 69 (1975).

[78] DOBBS ET AL., *supra* note 4, at 323; *see also* Richard W. Wright & Ingeborg Puppe, *Causation: Linguistic, Philosophical, Legal and Economic*, 91 CHI.-KENT L. REV. 461, 474, 480–81 (2016).

[79] As suggested in the Restatement (Third), the substantial-factor test "has proved confusing and been misused." RESTATEMENT (THIRD) OF TORTS: LIAB. FOR PHYSICAL AND EMOTIONAL HARM § 26 cmt. j (AM. L. INST. 2010).

[80] *See Nassar*, 570 U.S. at 347 (referring explicitly to tort law in determining the standard of causation applicable in a Title VII retaliation claim, and indicating that this "is the background against which Congress legislated in enacting Title VII, and these are the default rules it is presumed to have incorporated, absent an indication to the contrary in the statute itself").

[81] *See* Comcast Corp. v. Nat'l Ass'n of Afr. Am.-Owned Media, 140 S. Ct. 1009, 1013–14 (2020) ("Few legal principles are better established than the rule requiring a plaintiff to establish causation. In the law of torts, this usually means a plaintiff must first plead and then prove that its injury would not have occurred 'but for' the defendant's unlawful conduct.").

[82] *See generally id.* at 1017–18; *supra* Section II.A.

[83] *See infra* Sections III.A, V.A.

particular cases. As a consequence, standards of causation in disparate-treatment cases, like those in msc cases generally, have become increasingly jumbled, illogical, and ineffective.[84]

## III. A STRONGER FOUNDATION FOR THE CAUSAL INQUIRY

In this Part, I develop a theoretical foundation for the factorial framework and its applicability to disparate-treatment claims. I begin by discussing the importance of the necessity condition—the essential feature of the but-for test and a feature that is lacking in the motivating-factor test. I then introduce the potential-outcomes framework, a widely applicable counterfactual model of cause and effect in the sciences. I place the but-for test within this broader framework, and I show how the but-for test is only one measure of cause and effect within this broader counterfactual model. I then explain why other measures—and NESS in particular—retain the essential necessity condition, but in a broader form that is more appropriate for disparate-treatment claims and msc situations in general.

### *A. THE IMPORTANCE OF "MAKING A DIFFERENCE"*

Let us begin by asking what it means for a protected characteristic, such as race, to be a motivating factor. Consider circumstances in which a racist employer hires a white applicant over the plaintiff, an African-American applicant. Assume that, while the employer would favor the white applicant based on his race, he would hire the white applicant based on his qualifications alone, regardless of race. This is a typical mixed-motive scenario. But what makes race a motivating factor if it made no difference in the outcome of the employment decision?

One possibility is that the employer *considered* race in rendering his employment decision.[85] But precisely what does this mean? For example, is it necessary that the employer's consideration of race be conscious? A virulent racist who would never hire an African American over a white applicant may have no *conscious* operation or accounting of race. If *conscious* consideration of race is not necessary, however, is an employer's racism alone sufficient for race to qualify as a motivating factor? For example, if an African-American candidate applies for a position (perhaps among a dozen other applicants), and the employer is shown to be racist through the employer's statements and use of racial slurs two years earlier, is the employer, without more, liable for discrimination if he does not hire the African-American applicant?

We are interested in examining the meaning of the motivating-factor test in the absence of the necessity condition, the notion of "making a difference." Therefore, we cannot logically ask the obvious question: did the employer's racism make any difference in the hiring process? The motivating-factor test

[84] *See Counterfactual Causation*, *supra* note 15, at 879, 884–93.

[85] *See* Price Waterhouse v. Hopkins, 490 U.S. 228, 239–40 (1989).

does not depend on this. We can ask whether the employer's racism played a role in the hiring decision; but again, what does this mean in light of the facts above? Assume there is no evidence of any explicit or even conscious consideration of race in the hiring process. But the employer is a known racist, based on substantial evidence. Is the employer liable for discrimination if he hires a white applicant who happens to be more qualified for the position? What if, additionally, the company—say, a delivery service—is seeking to hire a driver, and by "more qualified," we mean that a white applicant has a driver's license whereas the African-American applicant does not. In other words, the African-American applicant is legally unable to complete the primary function of the position. Is the employer to be held liable for hiring a white applicant over the African-American applicant?

Next, let us ask whether our conclusion would change if the employer made a racial remark during the interview that made clear that he consciously considered race in his hiring decision, but where the only African-American candidate in the applicant pool was not licensed to drive. Even if so—if *conscious* consideration is the key—then a number of other problems arise. First, it is unclear why consciousness should be understood as the critical element since, by this measure, a virulent racist could hire white applicants over African-American applicants 100 out of 100 times with no liability for discrimination if he gives no *conscious* consideration to the race of the applicants. Remember, this standard does not account for "making a difference." Perhaps the proportion of African-American hires could be used as *evidence* of the employer's mental state, but it is not an aspect of the measurement of interest. Second, and relatedly, what does *conscious* consideration mean? The employer of course notices that an applicant is male or female, or that the applicant is African American, white, or another race, or that an applicant is wearing a yarmulke or a turban. Is this sufficient? Would it be sufficient for liability if, during the interview of a Jewish applicant, the employer thought of an anti-Semitic stereotype he had heard two years earlier, but quickly pushed it out of his mind and continued the interview? What if he thought about it for a short while before pushing it from his mind? Or, what if he kept it in his mind throughout the hiring process and ultimately rejected the Jewish candidate's application—but only because the applicant was unlicensed to drive and therefore, Jewish or not, could not be hired for a driving position.

Finally, what if a racist employer considers race in his hiring decision but in fact hires an African-American applicant over a white applicant, notwithstanding his race, based on the African-American applicant's credentials? Would the African-American applicant have a cause of action against the employer grounded in discrimination? After all, the motivating-factor test is not outcome-based, and the employer did consider race in deciding whether to hire the applicant. Of course, we could ask whether there was any harm. However, the same could be asked of the applicant who was not licensed to drive and therefore could not have been hired whether the employer discriminated based on race or not. Furthermore, race *was* a motivating factor in the hiring decision. It ultimately did not make a difference in the hiring decision, but it was a conscious consideration just as it was when the unlicensed driving applicant did not get hired; and, in that case, it did not make a difference either.

All of this is to say that the motivating-factor test is not a test of causation. Moreover, it is vague conceptually and analytically. Even if courts could apply it in an analytically meaningful way, it is likely to mean one thing to one jury and a different thing to a different jury.

The motivating-factor test, therefore, essentially asks the factfinder to decide based on policy, or the factfinder's general sense of responsibility, whether to hold the employer liable. At the same time, however, we provide factfinders with no direction regarding policy or criteria for responsibility. We do not indicate, for example, whether to use a theory of deterrence or fairness or compensation as a guide for determining whether to hold the employer responsible. Rather, although we refer to the motivating-factor test as a standard of causation, we simply ask the factfinder to arrive at a determination based on its intuition.

### *B. The But-For Standard as Over-Restrictive*

The discussion above highlights the inadequacy of the motivating-factor test. Unfortunately, as indicated above, the but-for test gives rise to substantial problems as well.[86] To be sure, the but-for test is at least a straightforward and analytical test that produces results that align with common sense in most cases. The problem is that the but-for test denies causation in a substantial category of discrimination claims—mixed-motive claims—in which common sense, court decisions, and good policy dictate that causation exists.

If, as in the example above, a racist employer hired a white applicant due to his race, but the white applicant was incidentally far better qualified than the competing African-American applicant, using the but-for standard, the plaintiff would be unable to establish a discrimination claim because he would be unable to show that he would have been hired absent the discrimination. To the contrary, because he was far less qualified than the white applicant, it is unlikely that the plaintiff would have been hired even in the absence of any discrimination.

Although, in this scenario, race is not a but-for cause of the adverse employment action, it is arguably a cause in the ordinary, or common, sense of the term, just as each fire in the concurrent two-fire problem is a cause of the lodge's destruction notwithstanding a second sufficient fire. Moreover, capturing the employer's discriminatory treatment as a cause in this scenario

[86] *See supra* Section II.B.

likely fulfills fairness and deterrence objectives of antidiscrimination law. This is well-supported by case law and scholarship regarding mixed-motive cases and msc situations generally.[87]

Thus, as a normative matter and not only as a descriptive matter, neither the motivating-factor test nor the traditional but-for test is adequate as a standard of causation. The motivating-factor test is not a coherent test of causation. Arguably, it provides little more guidance than instructing the jury, "do what you think." At the same time, the but-for test is overrestrictive and leads to the inappropriate exclusion of certain factors that common sense, court decisions, and good policy tell us are causes.

### *C. Potential Outcomes: A Robust Causal Framework*

The law need not choose between the motivating-factor test and the but-for test. Other good options are available. It is possible to apply a standard of causation that requires an employer's discriminatory conduct to have "made a difference" in the outcome of an employment decision without requiring a *but-for* difference. In particular, it is possible to employ a broader, more moderate form of "difference"—a less stringent version of the necessity condition. Applying this broader form of "difference" is not a compromise approach that manipulates the meaning of a "difference" and fabricates a new notion of cause and effect for purposes of achieving a sought-after outcome in mixed-motive cases and other msc situations. Instead, it reflects a well-accepted scientific model of cause and effect and should be understood as a *refinement* of both the motivating-factor test and the but-for test.

The factorial framework is based on the counterfactual model of cause and effect known as the potential-outcomes model. Understanding this broader scientific model is crucial for understanding the role of the but-for estimand and the appropriateness of a broader estimand in msc situations, as well as for attaining a refined theory and method of causation in disparate-treatment cases.

In law, the counterfactual model of causation, or "counterfactual causation,"[88] is frequently synonymized with but-for causation. But it is actually a broader concept. In the sciences, the potential-outcomes framework represents this broader counterfactual concept.[89] Its significance in law, more than a mere analogy, derives from the fact that factual causation is intended to reflect the actual, scientific, and common notion of cause and effect.[90] This

[87] *See infra* Part V.

[88] *See generally Counterfactual Causation*, *supra* note 15 (examining the counterfactual model of cause and effect in law and in the sciences).

[89] *Id.*

[90] *See* Richard W. Wright, *The NESS Account of Natural Causation: A Response to Criticisms*, *in* PERSPECTIVES ON CAUSATION 285, 285 (Richard Goldberg ed., 2011) ("natural (scientific, 'actual', 'factual') causation").

is clear from case law and scholarship, as well as from the regular application of statistical evidence of cause and effect as defined in the sciences as evidence of causation.[91] It is also supported by the close alignment between causal language in ordinary or common usage and the causal concept in the sciences.[92]

### 1. Identifying the Building Blocks of the Causal Inquiry

Under the potential-outcomes framework, causal inference starts with defining the building blocks of a causal effect and precisely specifying the causal estimand to be studied. The framework involves defining interventions, or "treatments," and conceptualizing and examining causal effects based on comparisons between "potential outcomes"—that is, outcomes that would be realized under different treatments.[93]

Consider, for example, a study aimed at determining the causal effect of pain medication with respect to a patient's backpain.[94] A researcher would begin by defining the "primitives" of the causal inquiry.[95] She would define a "unit," a person or object at a certain point in time; a "treatment," defined as "an action or intervention that can be initiated or withheld from that unit" at a certain point in time; an "outcome variable," a quantity of interest that is hypothesized to be affected by the treatment; and "potential outcomes," the outcomes (particular values of the specified outcome variable) that would be realized if one particular treatment or another is assigned to a particular unit.[96]

For example, the researcher may define a patient as a unit; define "active" and "control" treatments, or "treatment conditions," as the administration and non-administration of pain medication, respectively; and define the outcome variable as the patient's level of backpain.[97] She can then define

[91] *See infra* Section V.B; *see also Counterfactual Causation*, *supra* note 15, at 900–02 (noting that courts "accept and sometimes *require* (or at least hold as the 'gold standard') proof of causation through statistical evidence establishing a causal connection between a defendant's misconduct and a plaintiff's injury," and citing cases).

[92] *See infra* notes 233–43 and accompanying text.

[93] *See* GUIDO W. IMBENS & DONALD B. RUBIN, CAUSAL INFERENCE FOR STATISTICS, SOCIAL, AND BIOMEDICAL SCIENCES: AN INTRODUCTION 3–30 (2015). Throughout this Section, and elsewhere in this Article, I employ causal-inference terminology and notation based on the Rubin Causal Model. *See id.*

[94] *See id.* (considering an example involving the effect of aspirin on a headache).

[95] D. James Greiner & Donald B. Rubin, *Causal Effects of Perceived Immutable Characteristics*, 93 REV. ECON. & STAT. 775, 775–78 (2011); D. James Greiner, *Causal Inference in Civil Rights Litigation*, 122 HARV. L. REV. 533, 576–79 (2008).

[96] IMBENS & RUBIN, *supra* note 93, at 3–7; Donald B. Rubin, *For Objective Causal Inference, Design Trumps Analysis*, 2 ANNALS APPLIED STAT. 808, 811–13 (2008); *see also Counterfactual Causation*, *supra* note 15, at 894–95; Greiner, *supra* note 95, at 558–61.

[97] *See Counterfactual Causation*, *supra* note 15, at 894.

potential outcomes Y(medication) and Y(no medication) as potential pain outcomes that would occur if the patient receives medication and if the patient receives no medication, respectively. A causal effect can then be specified as a comparison between the potential outcomes Y(medication) and Y(no medication).[98] For example, assuming a binary outcome variable, we can determine a causal effect based on four possible comparisons: (1) Y(medication) = pain vs. Y(no medication) = pain; (2) Y(medication) = no pain vs. Y(no medication) = pain; (3) Y(medication) = pain vs. Y(no medication) = no pain; and (4) Y(medication) = no pain vs. Y(no medication) = no pain.[99] Using these comparisons of potential outcomes, the first and fourth comparisons indicate the absence of a causal effect while the second and third comparisons indicate the existence of a causal effect, with the second comparison indicating that medication reduces pain and the third comparison indicating that medication increases pain.[100]

Importantly, of the four possible outcomes—Y(medication) = pain, Y(medication) = no pain, Y(no medication) = pain, Y(no medication) = no pain—only one can be observed: the unit will receive only one treatment —medication or no medication—and will realize only one outcome—pain or no pain. If the patient received medication and realized an outcome of no pain, a researcher cannot know what the outcome would have been had the patient received no medication. For this reason, under the Rubin Causal Model, causal inference is described as a "*missing data problem*": "given any treatment assigned to an individual unit, the potential outcome associated with any alternate treatment is missing."[101]

Therefore, determining a causal effect requires *inference*. In particular, it requires inferring a missing potential outcome—the potential outcome associated with the counterfactual that did not occur (e.g., the treatment "no medication")—and comparing the inferred potential outcome to the *observed* potential outcome, the potential outcome associated with the treatment that occurred (e.g., "medication").[102]

Sometimes, a researcher may be interested in learning about the effects of multiple treatments, or "factors," as well as the interactions among them. A

[98] IMBENS & RUBIN, *supra* note 93, at 3–7; *see also Counterfactual Causation*, *supra* note 15, at 894–95.

[99] *See* IMBENS & RUBIN, *supra* note 93, at 3–7; *see also Counterfactual Causation*, *supra* note 15, at 894–95. We can similarly define the outcome variable, and therefore the potential outcomes, in terms of a pain rating from one to ten, and a causal estimand based on these potential outcomes. *See id.* at 895–97.

[100] *See* IMBENS & RUBIN, *supra* note 93, at 5–7; *see also Counterfactual Causation*, *supra* note 15, at 894–95.

[101] IMBENS & RUBIN, *supra* note 93, at 14. *See generally* Donald B. Rubin, *Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies*, 66 J. EDUC. PSYCH. 688 (1974).

[102] *See* IMBENS & RUBIN, *supra* note 93, at 3–30. For a more extended discussion of estimating causal effects, see *Counterfactual Causation*, *supra* note 15, at 896–97.

study with a "factorial" design is one in which each unit is exposed to a "treatment combination" (or "treatment condition") that is composed of a set of factors, where each factor can be set to a certain value, or "level."[103] For example, if the researcher above wanted to test the effect of the pain medication *and* physical therapy on the patient's backpain, she would define the pain medication and physical therapy as factors, each of which may take values "on" and "off." She could then define causal effects based on comparisons between various treatment conditions involving different combinations of these factors. She may, for example, be interested in studying the effect of medicine and therapy together as compared to neither medicine nor therapy, or as compared to only medicine or only therapy.[104]

### 2. Replication, Experimentation, and Covariate Balance

In statistics and the sciences, causal inference is facilitated by replication, and, in particular, by exposing multiple units to different treatments in an "experiment."[105] In an experiment, each unit is exposed to one treatment condition, and while we can observe the potential outcome associated with that treatment condition, we cannot observe the potential outcomes associated with treatment conditions not assigned to that unit. Replication, however, allows the researcher to impute missing potential outcomes and estimate causal effects.[106]

Frequently, experimentation—and particularly, controlling the assignment of treatments to units—is not possible. In such circumstances, scientists employ "observational studies," or studies in which the researcher does not control the assignment of treatments to units, but uses statistical methods to make causal inferences.[107] The Rubin Causal Model takes a prospective approach to observational studies: it generally seeks to "approximate, or attempt to replicate, a randomized experiment" by carefully (and objectively) defining causal estimands and by seeking to recreate comparisons that may result from a randomized experiment.[108] A causal inquiry in a legal case can be understood as a type of observational study

[103] Hillel J. Bavli & Reagan Mozer, *The Effects of Comparable-Case Guidance on Awards for Pain and Suffering and Punitive Damages: Evidence from a Randomized Controlled Trial*, 37 YALE L. & POL'Y REV. 405, 420 (2019); Tirthankar Dasgupta, Natesh S. Pillai & Donald B. Rubin, *Causal Inference from $2^K$ Factorial Designs by Using Potential Outcomes*, 77 J. ROYAL STAT. SOC'Y. SERIES B (STAT. METHODOLOGY) 727, 727 (2015) (proposing "[a] framework for causal inference from two-level factorial designs . . . which uses potential outcomes to define causal effects").

[104] *See generally* Dasgupta et al., *supra* note 103.

[105] IMBENS & RUBIN, *supra* note 93, at 3–30.

[106] *See Counterfactual Causation*, *supra* note 15, at 896.

[107] *See* IMBENS & RUBIN, *supra* note 93, at 41–42. *See generally* PAUL R. ROSENBAUM, OBSERVATION AND EXPERIMENT: AN INTRODUCTION TO CAUSAL INFERENCE (2017).

[108] Donald B. Rubin, *The Design Versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials*, 26 STAT. MED. 20, 25 (2007).

—one that involves inferring causal effects from evidence (statistical or other).

Central to causal inference is the problem that replication will always be imperfect since each unit exists only at a certain point in time. Assigning a patient medication today and no medication tomorrow to test the difference between backpain with medication and without medication involves imperfect replication. This experiment involves two units: the patient today and the patient tomorrow. There may be important differences between the patient today and tomorrow—differences other than whether he receives medication—and these differences, and not the treatment, could account for differences in his backpain.[109] Even assigning medication and no medication to identical twins at the same point in time would involve imperfect replication, since even identical twins have differences.[110]

Imperfect replication can cause poor imputation of missing potential outcomes and therefore poor inference regarding causal effects. The problem boils down to different units having different "covariates"—that is, a unit's background characteristics that are not affected by assignment to one treatment or another.[111] These can include age, race, sex, blood type, height, pretreatment measures (such as pretreatment income or a pretreatment level of backpain), and other characteristics. When units have different covariates, replication becomes less useful because a researcher will have more difficulty knowing whether observed differences between potential outcomes are attributable to differences in treatment (e.g., medication or no medication) or to differences in covariates (e.g., differences in pretreatment backpain of one unit versus another, or differences in pretreatment weight, which may impact the effectiveness of a medication).[112]

The primary advantage of an experimental study is that the researcher has control over the assignment of treatment conditions to units. Therefore, the researcher can, for example, assign treatment conditions to units randomly. "Randomized experiments" facilitate "balancing" covariates across treatment groups, so that a researcher can be more confident in her estimation of causal effects.[113] For example, if a researcher has a large sample of patients, and she randomly assigns the treatment conditions "medication" and "no medication" to them, it is likely that the pretreatment level of backpain (and other covariates) in the "medication" group and the "no medication" group will be approximately the same. Therefore, a difference in the observed outcomes of both groups with respect to backpain is more likely

[109] *Counterfactual Causation*, *supra* note 15, at 897–900.

[110] *Id.*

[111] *See* IMBENS & RUBIN, *supra* note 93, at 15–16.

[112] *Counterfactual Causation*, *supra* note 15, at 897–900.

[113] *See* Rubin, *supra* note 96, at 809–10.

to be attributable to the treatment rather than a difference in pretreatment backpain (or other covariates).

On the other hand, if a researcher is unable to control the assignment of treatments, she must be especially conscious of covariate balance among treatment groups.[114] For example, if she performs an observational study by comparing backpain outcomes in individuals who took backpain medication to backpain outcomes in individuals who did not take backpain medication without ensuring covariate balance, she is likely to obtain a misleading result. After all, individuals who take backpain medication generally have far worse backpain than those who do not take backpain medication. Even after taking medication, they are likely still to have worse backpain than, for example, the general public. If the researcher is not concerned with covariate balance, she may conclude that the backpain medication causes *more* backpain, whereas, in truth, the difference is likely attributable simply to the difference in pretreatment backpain. In other words, the researcher failed to compare apples to apples and therefore obtained a misleading result.

This type of error can sometimes be avoided with careful attention to account for relevant covariates. When this is not possible, good inference from this kind of data may not be possible: the data may be unreliable for drawing causal conclusions.

When a researcher is attentive to ensuring covariate balance, she must distinguish between covariates and another type of variable known as an "intermediate variable." Unlike a covariate, an intermediate variable can be affected by the treatment that a unit receives.[115] For example, if the side effect of a medicine is thirstiness, it would be a mistake to define thirstiness as a covariate and to balance thirstiness between the "medication" and "no medication" groups. Doing so may cause a misleading result of "no effect," for example, if the medicine's effectiveness is highly associated with thirstiness. In this case, ensuring that treatment groups had equal levels of thirstiness would obscure the effect of the medication.

An important consideration in determining whether a variable is a covariate or an intermediate variable is the timing of the treatment.[116] Remember that our definition of a treatment includes a time component, as does our definition of a unit.[117] In some studies, identifying the timing of treatment is simple. For example, in a laboratory experiment in which a researcher applies one medication or another to a patient, the operative time

[114] Note that a researcher must be vigilant with respect to covariate balance even if she is able to randomize treatment assignments. For example, a researcher should check covariate balance among treatment groups to confirm that the randomization was successful in achieving balance. *See, e.g.*, Bavli & Mozer, *supra* note 103, at 426–28.

[115] *See* Greiner, *supra* note 95, at 565–66.

[116] *See id.*

[117] *See supra* notes 94–96 and accompanying text.

for precisely defining the treatment and unit is the time of that application. As we will see below in the disparate-treatment context, however, determining the timing of a treatment can be more complex.[118]

### 3. Defining a Causal Effect

The potential-outcomes notion of cause and effect forms the basis (implicitly or explicitly) for the but-for concept in law and in common usage. However, the but-for standard is only a particular measure of cause and effect under the potential-outcomes framework. To describe a broader set of effects, let us consider the simple example of the two-fire problem, and, in particular, the causal effect of Fire A, within the context of a factorial design in the potential-outcomes framework.[119] We define two factors—Fire A and Fire B—and an outcome variable of interest—a binary variable representing whether the lodge in the two-fire problem is destroyed or not destroyed. We can then consider potential outcomes associated with each combination of levels—"off" and "on"—linked to each factor. Figure 1 illustrates the potential outcomes that describe a concurrent msc situation involving Fire A and Fire B.

Figure 1. 2x2 matrix illustrating potential outcomes associated with combinations of two factors, Fire A and Fire B, each with two levels, off and on

| | **Fire A = off** | **Fire A = on** |
|---|---|---|
| **Fire B = off** | Not Destroyed | Destroyed |
| **Fire B = on** | Destroyed | Destroyed |

As assumed in the two-fire concurrent msc problem, the lodge is destroyed when either Fire A or Fire B is on (upper right and lower left quadrants) and when both Fire A and Fire B are on (lower right quadrant), but not when neither Fire A nor Fire B is on (upper left quadrant).[120]

Defining a causal effect involves specifying a comparison between potential outcomes associated with different levels of a factor. For example, defining a causal effect of Fire A on the lodge's destruction involves a comparison between potential outcomes associated with the two levels ("off" and "on") of Fire A. Let us distinguish, however, between two categories of

[118] *See infra* Part IV; Greiner & Rubin, *supra* note 95, at 775–78; Greiner, *supra* note 95, at 576–79.

[119] *See generally* Dasgupta et al., *supra* note 103.

[120] *See Counterfactual Causation, supra* note 15, at 907–08.

causal effects in this scenario—"unconditional" effects and "conditional" effects—based on whether the comparison of interest is unconditional or conditional on the level of another factor. We can refer to these effects collectively as "main effects."[121]

Let me pause here to make two important notes regarding terminology. First, in this Section's discussion, I simplify matters by using a scenario (the two-fire scenario) that involves two factors, each having two levels, and by classifying effects as either "unconditional" or "conditional." Indeed, this simple model is applicable to many real-world scenarios. However, in more complex scenarios involving additional factors, the unconditional/conditional classification can easily be extended to describe various levels of conditioning. Additionally, I use the term "unconditional" here to refer in particular to the absence of any conditioning on the level of the second factor (e.g., conditioning on Fire B being "on"). Second, I use the term "main effects" (or a "main-effects analysis") to capture both unconditional and conditional effects, *but with special emphasis on unconditional effects.* I contrast it with the but-for test, emphasizing the breadth of a main-effects analysis and its ability to capture effects that the but-for test excludes. I sometimes use the term "unconditional effects" or "unconditional main effects" to emphasize that the effect to which I am referring does not involve conditioning on a particular level of a second factor. However, consistent with scientific literature, I often use the term "main effects" (or a "main-effects analysis") to refer to *unconditional* effects in particular. On the other hand, when referring in particular to effects involving conditioning on the level of a second factor, I always use the term "conditional effects" or "conditional main effects."

Now, in a single-factor scenario, the but-for standard reflects both types of effects. But this is not necessarily so in multifactor situations. In the two-fire problem illustrated in Figure 1, when considering the causal effect of Fire A, the but-for test takes Fire B = on as given and ignores the potential outcomes associated with Fire B = off. It therefore compares the potential outcomes associated with the combinations (Fire A = on, Fire B = on) and (Fire A = off, Fire B = on). That is, it compares the potential outcome in the lower right quadrant with the potential outcome in the lower left quadrant. Because the lodge is destroyed for each of these combinations, the but-for test leads to the conclusion that there is no causal effect.

This causal question involves a *conditional* effect: it is based on a comparison between potential outcomes associated with two levels of a factor (Fire A = on versus Fire A = off) while holding the level of another factor (Fire B = on) constant.

[121] *See generally* Dasgupta et al., *supra* note 103. In this Article, I reserve the term "interaction effects," *see Counterfactual Causation, supra* note 15, at 906–08, to refer to contrasts between conditional main effects. *See* Dasgupta et al., *supra* note 103, at 730.

Using the numerical values 1 and 0 to represent "destroyed" and "not destroyed," the conditional effect of Fire A on the destruction of the lodge when Fire B = on is 1 – 1 = 0. That is, the but-for effect of Fire A is zero, or nothing, since, in the absence of Fire A—given the existence of Fire B—the lodge would have been destroyed anyway.[122] Similarly, the conditional effect of Fire B when Fire A = on is zero. There is no effect.

Unconditional main effects are different than conditional main effects. The former effect is based on a comparison between potential outcomes associated with two levels of a factor *without* holding the level of another factor constant. Therefore, the unconditional main effect of Fire A on the state of the lodge compares both potential outcomes associated with Fire A = on to both potential outcomes associated with Fire A = off. There are various ways to conduct such a comparison. For example, the effect could be defined as the difference between the *average* of the two potential outcomes associated with Fire A = on and the *average* of the two potential outcomes associated with Fire A = off, where each average combines (e.g., using the mean) the two potential outcomes associated with each level of Fire A across each level of Fire B. Using this definition (for illustrative purposes only), the effect of Fire A, using the notation above, would be (1 + 1)/2 – (0 + 1)/2 = 0.5. Therefore, unlike our determination using the but-for standard, our unconditional main-effects analysis results in a positive effect, and findings of causation for both Fire A and Fire B.[123]

The hallmark feature of an unconditional main-effects analysis is that it defines a causal effect based on the potential outcomes associated with the full range of treatment levels. In our example, it considers not only counterfactuals associated with Fire A given the existence of Fire B but also counterfactuals associated with Fire B. In other words, it asks not only what would have happened had Fire A not occurred, given that Fire B occurred, but also what would have happened had Fire A occurred versus not occurred if Fire B had *not occurred*. In the following Section, I explain why this broader main-effects estimand is appropriate for msc situations.

### *D. The Logic of a Main-Effects Analysis—A Broader Form of "Difference"*

It would not be logical to rule out an unconditional main-effects analysis on grounds that it involves consideration of counterfactuals associated with the non-occurrence of Fire B when we know for certain (by definition of the two-fire msc problem) that Fire B in fact occurred. This is because the very essence of the counterfactual model, and the but-for test in particular, is the comparison of counterfactuals—the inference of cause and effect based not only on what actually occurred but also what would have occurred, but did not, under counterfactual circumstances. An unconditional main-effects

[122] *See Counterfactual Causation*, *supra* note 15, at 907–08.

[123] *See id.*

analysis in the two-fire problem simply considers counterfactuals associated with both Fire A *and* Fire B.

Nevertheless, a valid question remains: even if it is logical not to rule out an unconditional main-effects analysis, why does it make sense to employ this analysis rather than a *conditional* main-effects analysis, and the but-for test in particular, if we indeed know that Fire B occurred? The answer is tied directly to the near-universal consensus that the but-for test fails in msc situations.[124]

The but-for test is said to fail in msc situations because its outcome in these situations is contrary to common sense and ordinary usage and thinking regarding the notion of cause and effect.[125] However, the counterintuitive result arises from the law's misapplication of the counterfactual model rather than the failure of the model.[126] In particular, the confusion arises from the failure to adjust the traditional but-for standard of causation to account for a multifactor setting.

The reason it is deemed incorrect to conclude that Fire A in the two-fire problem is not a cause of the lodge's destruction is this: *Had Fire B not occurred, Fire A would have destroyed the lodge; therefore, it is illogical to conclude that, simply because Fire B also occurred, Fire A is not a cause.*[127] This reasoning, however, follows an unconditional main-effect analysis, which defines the causal effect of Fire A on the state of the lodge based not only on what would have happened had Fire A not occurred (i.e., the potential outcomes associated with Fire A = on versus Fire A = off, given the occurrence of Fire B) but also what would have happened had Fire B not occurred (i.e., the potential outcomes associated with Fire A = on versus Fire A = off, given the non-occurrence of Fire B). In other words, the unconditional main-effects analysis asks exactly what is stated in the italicized reasoning above: The conclusion that Fire A is not a cause is illogical because had Fire B not occurred, Fire A would have destroyed the lodge—referring explicitly to the potential outcomes associated with the *non-occurrence* of Fire B, as well as those associated with the occurrence of Fire B.[128]

Thus, the law's application of a conditional main-effects analysis rather than an unconditional main-effects analysis can, and often does, lead to a counterintuitive result in msc situations. Our intuition regarding cause and effect often employs an unconditional main-effects analysis. There is no reason that the law's standard of factual causation—intended to capture the

[124] *See supra* notes 60–66 and accompanying text.

[125] *See supra* notes 60–66 and accompanying text.

[126] *Counterfactual Causation*, *supra* note 15, at 907–08.

[127] *Id.* at 911–12.

[128] *Id.*

common and scientific meaning of cause and effect—should not reflect this.[129]

### *E. NESS: A (LIBERAL) MAIN-EFFECTS MEASURE*

The NESS test can be used to define a causal effect in an unconditional main-effects analysis. As explained in *Counterfactual Causation*, an unconditional main-effects analysis "can be simplified and applied in practice by using the causal-set approach of the NESS test and the Restatement Third [of Torts]."[130]

The NESS test is based on the concept of a causal set—a set of factors that together lead to the occurrence of an outcome.[131] It asks whether a factor was a necessary element of a sufficient set.[132] According to this test, "a condition contributed to some consequence if and only if it was necessary for the sufficiency of a set of existing antecedent conditions that was sufficient for the occurrence of the consequence."[133] For example, the fires in the two-fire problem are NESS causes because each fire (individually) forms a set (of antecedent conditions) that was sufficient for the destruction of the lodge, where each was necessary for the sufficiency of the set. In a more complex example involving three fires, where any two of the fires would have been sufficient to destroy the lodge, each fire is a cause under the NESS test because each fire is part of a set of two fires that were sufficient to bring about the destruction of the lodge, where each fire was necessary for the sufficiency of the set in producing the lodge's destruction.[134]

A simple version of this approach has been adopted by the Restatement (Third) of Torts as the standard of causation in msc situations. In contrast with the Restatement (Second) of Torts, which uses the substantial-factor test as its primary definition of causation,[135] section 26 of the Restatement (Third) defines "factual causation" as a cause if it satisfies the but-for test: "Tortious conduct must be a factual cause of harm for liability to be imposed. Conduct is a factual cause of harm when the harm would not have occurred absent the conduct. Tortious conduct may also be a factual cause of harm under § 27."[136] Section 27 of the Restatement (Third) provides for msc situations: "If multiple

[129] *Id.*

[130] *Id.* at 917.

[131] *See supra* note 16 and accompanying text.

[132] *See* Wright, *supra* note 55, at 1740.

[133] Wright, *Grounds and Extent*, *supra* note 17, at 1441; Wright, *Once More into the Bramble Bush*, *supra* note 17, at 1102–03.

[134] *See* RESTATEMENT (THIRD) OF TORTS: LIAB. FOR PHYSICAL & EMOTIONAL HARM § 27 cmt. f illus. 3 (AM. L. INST. 2010).

[135] RESTATEMENT (SECOND) OF TORTS § 431 (AM. L. INST. 1965).

[136] RESTATEMENT (THIRD) OF TORTS: LIAB. FOR PHYSICAL & EMOTIONAL HARM § 26 (AM. L. INST. 2010).

acts occur, each of which under § 26 alone would have been a factual cause of the physical harm at the same time in the absence of the other act(s), each act is regarded as a factual cause of the harm."[137]

Applying the Restatement (Third) version of the test to the two-fire problem, had each fire occurred alone (i.e., in the absence of the other fire), each individually would have constituted a but-for cause of the lodge's destruction; therefore, each fire is a cause under section 27. The Restatement (Third) approach is intended to apply similarly to the three-fire problem described above, with all three fires constituting causes under this approach.[138]

As discussed above, the factorial framework allows for various measures, or "estimands," that a court could employ as a standard of causation. The traditional but-for test is a particular conditional effect. However, as discussed, it may be more appropriate to employ a broader main-effects analysis in msc situations.

The NESS test and the derivative test employed by the Restatement (Third) can be understood as a particular unconditional main-effects estimand.[139] Let us consider the two-fire problem to understand why this is. An unconditional main-effects analysis compares the potential outcomes associated with Fire A = on to the potential outcomes associated with Fire A = off without limiting the comparison to a particular level of Fire B. It aggregates the potential outcomes in the third column of Figure 2 (the potential outcomes associated with Fire A = on) and separately aggregates the potential outcomes in the second column (the potential outcomes associated with Fire A = off) and then compares the two aggregations (e.g., using subtraction or division).

Figure 2. 2x2 matrix illustrating potential outcomes for combinations of two factors, Fire A and Fire B, each with two levels, on and off[140]

| | **Fire A = off** | **Fire A = on** |
|---|---|---|
| **Fire B = off** | Not Destroyed | Destroyed |
| **Fire B = on** | Destroyed | Destroyed |

[137] *Id.* § 27.

[138] *See id.* § 27 cmt. f illus. 3.

[139] *Counterfactual Causation*, *supra* note 15, at 920.

[140] *Id.* at 907–08.

There are numerous ways in which the sets of potential outcomes could be aggregated and then compared. One example, discussed briefly above, would be to average the potential outcomes associated with Fire A = on, average the potential outcomes associated with Fire A = off, and then subtract the latter from the former to determine the effect.[141] To find an effect, a main-effects analysis looks for differences between the two sets of potential outcomes. The particular main-effects estimand determines how such differences translate into the "causal effect."

Using the averaging method above, if it is assumed that a studied factor only affects the outcome in one direction—for example, that pain medication in a particular scenario can make pain better but not worse, that pollutants in a particular scenario can make the air more polluted but not less polluted, or that discrimination against African Americans in a particular scenario can make it more difficult for an African American to be hired but not less difficult—then this method is essentially equivalent to the approach of the NESS test and the Restatement (Third).[142]

More broadly, the approach of the NESS test and the Restatement (Third) can be understood as involving a broad unconditional main-effects estimand that looks for the minimum difference between sets of potential outcomes associated with different levels of a factor, such as the occurrence and non-occurrence of a fire. It results in a determination of "causation" if there is, for example, any difference between matrix columns reflecting the occurrence and non-occurrence of a factor.[143] Fire A is therefore a cause of the lodge's destruction because, either when Fire B = off or when Fire B = on, Fire A being on versus off makes a difference. In particular, in the 2x2 matrix in Figure 2, when Fire B = off, the lodge is destroyed when Fire A = on, but not when Fire A = off. This explanation applies similarly to more complex msc situations.[144]

## IV. A FACTORIAL FRAMEWORK FOR DISPARATE-TREATMENT CASES

The factorial framework should be understood as both a general theory of causation for disparate-treatment cases and a concrete method for resolving causal inquiries in these cases. It is grounded in the potential-outcomes framework and the logic of main-effects estimands in msc situations. An important component of this approach is carefully defining the primitives of the causal inquiry, including units, treatments, and outcome variables, as well as the causal estimand—the measure of causation in terms

[141] *See supra* Section III.C.3.

[142] *See Counterfactual Causation*, *supra* note 15, at 920.

[143] *See id.* at 922 n.119 and accompanying text.

[144] *See id.* at 920–23.

of potential outcomes.[145] Understanding these concepts within a case enables a proper conceptualization of the causal inquiry and the evidentiary inferences required to make causal conclusions.

In this Part, I outline the proposed method for addressing causal questions in disparate-treatment claims. I also discuss the central role of the potential-outcomes framework, and the importance of "preemption" and the damages phase for preventing false findings of causation and windfall recoveries.

### A. *Proving Causation*

The factorial approach adopts the NESS test as the appropriate causal estimand in disparate-treatment claims—the standard of causation that must be satisfied to prove a claim. Applying the potential-outcomes model and the NESS test as its causal estimand allows for a simpler, more coherent, and more effective proof scheme.

A summary of this approach is as follows:

> 1. To establish a prima facie disparate-treatment claim, a plaintiff must satisfy a form of the *McDonnell Douglas* criteria: (a) membership in a protected class; (b) an open employment opportunity controlled by the defendant; (c) minimal qualification for the employment opportunity; (d) an adverse employment action; and (e) a causal link—using NESS causation and the potential-outcomes model —between the employer's discriminatory conduct and its adverse employment action.[146]
>
> 2. Although the plaintiff retains the burden of production and persuasion, the defendant may respond to the plaintiff's allegations by (a) rebutting the plaintiff's allegations, or (b) establishing a sufficient legitimate purpose.
>
> 3. If the defendant establishes a sufficient legitimate purpose, the plaintiff must satisfy her burdens of production and persuasion by establishing (a) that the protected feature is a NESS cause (even if not a but-for cause) of the adverse employment action, or (b) that the defendant's alleged legitimate purpose cannot explain the adverse employment action—either because the purpose is not applicable to the plaintiff or because the purpose was not in fact sufficient.

Let us consider this framework in greater detail. Consider a sex-discrimination claim filed by Monica, a female employee of a large technology

[145] Greiner & Rubin, *supra* note 95, at 775–78; Greiner, *supra* note 95, at 576–79; *see supra* Section III.C.

[146] *See generally supra* notes 26–28 and accompanying text.

corporation, alleging that her employer deprived her of a promotion due to her sex.[147] The employer acknowledged promoting Jerry, a male employee, over Monica, but argued that Jerry received the promotion because he was more qualified for the position.

Monica must begin by alleging facts sufficient to establish: (a) membership in a protected class; (b) an open employment opportunity controlled by the defendant; (c) minimal qualification for the employment opportunity; (d) an adverse employment action; and (e) a causal link between the employer's discriminatory conduct and its adverse employment action.[148] These elements

[147] I use the term "sex," as distinct from "gender," for consistency with the language of Title VII.

[148] Professor Brian Clarke has proposed a "revised *McDonnell Douglas* proof scheme" in which a plaintiff establishes a prima facie case by proving that "(1) the plaintiff has a protected trait; (2) the plaintiff was performing her job at a level that met the employer's reasonable expectations; (3) the plaintiff suffered an adverse employment action; and (4) the adverse action occurred under circumstances giving rise to an inference of discrimination." Brian S. Clarke, *A Better Route Through the Swamp: Causal Coherence in Disparate Treatment Doctrine*, 65 RUTGERS L. REV. 723, 780–81 (2013). Clarke's proposal involves a burden-shifting scheme in which a plaintiff's proof of her prima facie case would shift the burden to the defendant to "articulate *all* of the legitimate non-discriminatory reasons . . . for its decision." *Id.* "If the defendant satisfies this burden of production, the presumption of discrimination is rebutted" and "the plaintiff must then prove that the employer's consideration of her protected trait was a necessary element of the set of facts and circumstances that led to the defendant's decision." *Id.* at 781. The plaintiff, under Clarke's method, can do this by "either (1) showing that *at least one* of the defendant's stated [legitimate purposes] is unworthy of credence"—in which case it should, per Clarke's approach, be inferred that the employer "lied . . . to cover up its consideration of the employee's protected trait"—"or (2) presenting other evidence from which a reasonable jury could conclude that the employer's consideration of her protected trait was a necessary element of the set of facts and circumstances that led to the defendant's decision." *Id.* at 781–82.

The factorial framework differs from Clarke's proposal in various critical respects. First, it is premised on a theory of cause and effect grounded in the potential-outcomes model. This conceptual model is fundamental to both the theoretical foundation and the practical method proposed in this Article. Second, although both the factorial framework and Clarke's proposal seem to outwardly employ the NESS test, the causal standard in Clarke's proposal seems substantially narrower—as permitting a finding of causation when a "protected trait was a necessary element of the set of facts and circumstances that led to the defendant's decision." *Id.* at 781. But, because this test seems to capture only necessary elements of *the* set of elements *that (in fact)* "led to the defendant's decision," it is a version of the traditional but-for test. Compare this to the framework proposed herein, which requires only that the discriminatory force be a necessary element of *a set* of elements *sufficient for the occurrence* of the adverse decision. This estimand deviates substantially from the but-for test and constitutes a main-effects analysis in the potential-outcomes framework. In other words, a key feature of the NESS test, as employed in the factorial framework, is its ability to capture as causes forces that were not in fact necessary for the occurrence of the adverse outcome. Instead, it only requires fulfillment of the necessity condition in a broader sense—for example, that even absent certain other forces, a protected feature would have made a difference. Third, the factorial framework, in contrast with Clarke's proposal, incorporates the concepts of treatment timing, covariates, and "preemption," all crucial to its application of NESS, and to preventing unjustified findings of liability and windfall recoveries. Clarke's application of "unclean hands as an equitable affirmative defense to [certain] remedies" to prevent windfall recoveries is arguably insufficient in various respects—for example, as addressing only a certain type of windfall recovery and as allowing unjustified findings of liability.

are a form of those articulated in *McDonnell Douglas* to establish a prima facie discrimination claim.[149] In establishing these elements, a plaintiff must —implicitly or explicitly—identify the primitives of the causal inquiry. In this case, the primitives can be derived straightforwardly from Monica's claim: Monica is the sole unit; sex is the protected feature; the employer's perception of Monica's sex is the treatment variable; and Monica's promotion status is the outcome variable of interest.[150] The timing of the treatment can be defined as beginning the moment that the employer first perceived Monica's sex.[151]

Note that there are various reasons for basing a treatment variable, both in terms of timing and substance, on the *perception* of a plaintiff's sex, race, or other protected characteristic.[152] Some of these reasons are technical—for example, grounded in the meaning of a treatment and other causal concepts in the potential-outcomes framework—and others are substantive.[153] For our purposes, using perception as a basis for the timing of a treatment generally makes most sense with respect to the law's conception of discrimination and for understanding important concepts in the causal framework, such as preemption, covariate values, and intermediate variables.[154] Relatedly,

*Id.* at 785. Fourth, the factorial framework's foundation for reshaping the causal standard in disparate-treatment claims, in contrast with Clarke's proposal, is grounded in the common and scientific notions of cause and effect, the relationship between but-for causation and the broader counterfactual model, and a range of policy objectives, including deterrence, fairness, and efficiency. Fifth, the factorial framework seeks to pivot the proposed causal standard to move disparate-treatment proof schemes in the direction of the more basic structure applied in torts cases generally.

Clarke's proposal to employ NESS and a modified *McDonnell Douglas* test is innovative and sound. Indeed, these components are central features of the factorial framework. Clarke's proposal is a move in the right direction; however, for the foregoing reasons and various others, it seems to fall short of a solution to the current problems surrounding standards of causation in disparate-treatment cases.

[149] *See* McDonnell Douglas Corp. v. Green, 411 U.S. 792, 802 (1973); *see also supra* notes 26–35 and accompanying text.

[150] *See supra* notes 94–100.

[151] *See infra* Section IV.B; Greiner & Rubin, *supra* note 95, at 775–78; Greiner, *supra* note 95, at 576–79.

[152] Greiner, *supra* note 95, at 576–77.

[153] *Id.*; Greiner & Rubin, *supra* note 95, at 775.

[154] Greiner, *supra* note 95, at 576–77. As Professor Greiner has explained,

> The law's focus on the specific actor's decisionmaking requires, indeed compels, the analyst to regard as "given" the characteristics of individuals that were in place prior to the individuals' interaction with the actor. The only way to do that is to define the treatment as taking place at some moment of perception by the actor of the characteristic common to the group.

*Id.* Arguably, there are theoretical difficulties with defining a treatment variable in an observational study based directly on an "immutable characteristic, such as race or sex." Greiner & Rubin, *supra* note 95, at 775. This is due, in part, to "the impossibility of manipulating such traits in a way analogous to administering a treatment in a randomized experiment and the throughout this Article, references to discrimination, discriminatory factors, and protected characteristics as treatment variables should be interpreted in terms of the employer's *perception* of the employee's protected characteristic.

Thus, to establish a prima facie discrimination claim, Monica must establish causation based on the primitives of the causal inquiry and the NESS estimand. Although the elements of the claim are similar to those in the *McDonnell Douglas* test, the proposed approach requires an explicit showing of causation (with direct or circumstantial evidence), whereas some articulations of a prima facie discrimination claim under *McDonnell Douglas* do not.[155] On the other hand, at this early stage of the litigation, requiring the claimant to meet a standard of NESS causation should not erect a substantial barrier to establishing a claim as compared to the motivating-factor test. A showing of but-for causation would establish NESS causation since the former is a particular category of the latter. However, although NESS requires that the protected feature "made a difference," it is far less stringent than the but-for test. In practice—certainly at this early stage—it is more similar, in terms of proof, to the motivating-factor test. Critically, it may allow a plaintiff to avoid summary judgment even when she is unable to establish but-for causation. Indeed, the NESS test can be understood as a refinement of the motivating-factor test, and other than establishing elements (b) and (c) above, the plaintiff need not rule out alternative explanations for an adverse employment action, since even other sufficient factors would not preclude a finding of causation under the NESS test if, for example, in the absence of another factor, the protected feature would have made a difference.

Requiring a plaintiff to establish but-for causation at this point, as part of her prima facie claim, would trivialize the claim in the sense that a plaintiff could not be expected to foresee all possible legitimate factors that a defendant will raise. On the other hand, requiring a plaintiff to establish NESS causation as part of her prima facie claim does not require the plaintiff to foresee the defendant's argument; it just requires some evidence of a causal link, even if that link, for example, assumes the absence of legitimate factors.

A defendant may respond to a plaintiff's allegations in two ways. First, the defendant may dispute the plaintiff's allegations. For example, Monica's

danger of posttreatment bias stemming from the fact that almost all variables on which a researcher would like to condition are determined after an individual's conception." *Id.* (citing sources). For an extensive discussion of this issue, see *id.*

[155] Note that at the pleading stage, the proposed approach requires that a discrimination plaintiff allege explicitly a causal link between the discriminatory conduct and the adverse employment decision. A thorough examination of the pleading requirements is beyond the scope of this Article. However, in general, to avoid dismissal, the complaint must allege facts that, if "taken as true," "state a claim to relief that is plausible on its face"—in particular, as in other torts claims, the plaintiff must plausibly allege the elements of misconduct, causation, and harm. Bell Atl. Corp. v. Twombly, 550 U.S. 544, 556, 570 (2007).

employer may show why Monica was ineligible for a promotion or show that there was no promotion opportunity in the first instance. More generally, one important way for the defendant to respond to the plaintiff's allegations is by showing that any discrimination was (to use NESS terminology) "preempted" by a preexisting condition that was determinative of the outcome—i.e., that was sufficient for the occurrence of the alleged adverse employment action.[156]

As discussed in Section II.B, a preemptive-causation problem is a type of msc situation in which one force precedes the other, such as when one fire in the two-fire problem arrives to destroy a lodge immediately before the other fire arrives.[157] In these situations, the first force *and not the second* is deemed a cause of the outcome. This is true legally, intuitively, by definition of NESS causation, and in the potential-outcomes framework, all of which require that a force act on the unit prior to the determination of the outcome in order for it to qualify as a cause. In terms of the potential-outcomes framework, a treatment must precede an outcome,[158] and, in terms of the NESS test, a force must be *antecedent to* the consequence at issue.[159] In short, a showing that a set of conditions preempted the alleged discrimination—including the absence of an open employment opportunity or the plaintiff's failure to minimally qualify for the opportunity—amounts to proof that discrimination cannot be a cause of the adverse employment action. I return to this issue in Section IV.B.2.

Second, the employer may demonstrate facts establishing a sufficient legitimate purpose (or sufficient legitimate purposes) for the adverse employment action. By making such a showing, the employer establishes that discrimination is not a but-for cause of the adverse decision. Importantly, however, the legitimate purpose must itself be sufficient to bring about the decision, and not sufficient only in combination with the discriminatory purpose, since, otherwise, the legitimate purpose cannot rebut the plaintiff's allegations of a causal link.[160]

It is important to realize that a sufficient legitimate purpose does not preclude NESS causation, the standard required for the plaintiff to prove her claim. It does, however, preclude but-for causation, which, at least initially —by default, since the defendant has not yet introduced legitimate

[156] *See supra* Section II.B; Wright, *supra* note 55, at 1794–98 (explaining preemptive causation).

[157] *See supra* Section II.B; Wright, *supra* note 55, at 1794–98.

[158] *See Counterfactual Causation*, *supra* note 15, at 914–15; Greiner, *supra* note 95, at 558–60.

[159] *See* Wright, *supra* note 55, at 1774, 1795 ("[T]he NESS test states that a particular condition was a cause of a specific consequence if and only if it was a necessary element of a set of antecedent actual conditions that was sufficient for the occurrence of the consequence.").

[160] If the employer simply shows that a legitimate purpose was part of the set of forces that brought about the adverse action, this does not rule out but-for causation. For example, if an employer fires an employee due to a combination of the employee being African American and frequently arriving late to work, the employee's race is a but-for cause of the adverse employment action.

purposes—will often be the form of NESS causation that a plaintiff establishes in order to satisfy the elements of her prima facie claim.[161] Moreover, a showing of sufficient legitimate purposes specifies a concrete set of factors to consider in determining whether discrimination is a NESS cause of the adverse employment action, and the defendant's legitimate-purpose defense can in fact rebut NESS causation evidentiarily.[162]

Finally, although a plaintiff may have initially alleged NESS causation in a single-factor situation, if the defendant has established a sufficient legitimate purpose, the plaintiff may respond to the defendant's rebuttal in two ways. First, the plaintiff may produce evidence establishing that discrimination, although not a but-for cause, is a NESS cause of the adverse employment action.[163] In many cases, this showing will take the form articulated in the Restatement (Third) of Torts: discrimination will be deemed a cause if, in the absence of the legitimate factor, the protected feature would have made a difference.[164] For example, the test would ask whether, assuming Monica were equally qualified for the promotion, Monica's sex would mean the difference between getting and not getting the promotion. This test effectively asks the factfinder to compare the potential outcomes in the top row of the 2x2 matrix in Figure 3 in addition to the potential outcomes in the bottom row. If either of the rows entails a contrast of potential outcomes, comparing Male to Female, then the standard for causation is satisfied.

[161] Recall that establishing a force as a but-for cause necessarily establishes that it is a NESS cause also.

[162] Even if it is assumed that the plaintiff introduced evidence sufficient to establish a *high probability* of NESS causation as part of her prima facie case and prior to the defendant introducing a legitimate purpose, an inference of discrimination and causation may be heavily based on the absence of an alternative explanation for the adverse employment action. If, for example, the defendant introduces overwhelming evidence that he fired the plaintiff because she hit a customer, the question would then become whether the employer would have fired the plaintiff even had she not hit the customer. However, the plaintiff's earlier evidence of NESS causation, and the inference that it permitted, may now be substantially weaker in light of the defendant's legitimate-purpose defense. *See generally* Edward K. Cheng, *Reconceptualizing the Burden of Proof*, 122 YALE L.J. 1254 (2013).

[163] *See generally supra* note 162.

[164] *See supra* notes 136–38.

Figure 3. 2x2 matrix depicting potential outcomes in a msc situation in which the NESS estimand asks factfinders to compare potential outcomes associated with sex = Male to those associated with sex = Female for both "Strongly Qualified" and "Weakly Qualified," and to conclude that sex is a cause if either comparison yields a difference in potential outcomes

| | **Male** | **Female** |
|---|---|---|
| **Strongly Qualified** | Promoted | Not Promoted |
| **Weakly Qualified** | Not Promoted | Not Promoted |

This standard is far less demanding than but-for causation. For example, Monica may well be able to provide sufficient evidence to establish NESS causation even if the defendant provides strong evidence that Monica was less qualified for the promotion than her male colleague and that she would not have received the promotion regardless of whether she was female or male.

The second way in which the plaintiff may respond to the defendant's rebuttal is by showing that the legitimate purpose alleged cannot explain the adverse employment decision, either because the purpose is not applicable to the plaintiff or because the purpose was not in fact sufficient. This showing serves as indirect proof of causation because it demonstrates that the defendant's legitimate purpose can at most be sufficient only in combination with the discriminatory purpose and therefore cannot serve as a rebuttal to the plaintiff's prima facie claim.[165] This method of responding to a defendant's legitimate purpose is similar to showing that an alleged legitimate purpose is pretextual under current antidiscrimination proof schemes. However, it provides a concrete method for establishing the invalidity of a defendant's alleged legitimate purpose. If the legitimate purpose is not applicable or is not sufficient for the occurrence of the adverse employment decision, then it cannot explain the decision and is therefore invalid as a rebuttal of the plaintiff's prima facie discrimination claim—it is, in other words, pretextual.

Thus, both of the foregoing methods can be used to invalidate the defendant's legitimate-purpose rebuttal. A showing of insufficiency demonstrates that the alleged reason cannot serve as an explanation for the decision, and a showing that the discriminatory factor is a NESS cause of the decision establishes that, regardless of whether the defendant's legitimate

[165] Note that, in the factorial model, the issue of "pretext," in a certain sense, becomes somewhat less important, since proof of NESS causation is consistent with proof of a sufficient legitimate factor.

purpose is pretextual or not, it is at best one of *two* causes, with the other cause being illegitimate.

Note that throughout the discrimination case, the plaintiff retains the burdens of production and persuasion. Additionally, in contrast with Title VII's current proof scheme—which involves both the motivating-factor test and the but-for test—the proposed approach involves only a single causal estimand—NESS. Moreover, the proposed approach involves no burden shifting and no meaningful distinction between direct and indirect forms of evidence, except as these categories of evidence are distinct in the general torts context.[166] A plaintiff must prove NESS causation in order to prove her claim. A defendant may, as in other torts claims, produce evidence to rebut this element of the claim. If, in light of the defendant's evidence, no reasonable jury could conclude that discrimination was the cause of the adverse employment decision, then the plaintiff has not met her burden of production, and the court may grant judgment to the defendant. However, unlike current standards, which frequently employ the stringent but-for estimand, the plaintiff must only produce sufficient evidence for a reasonable jury to find that the employer's discrimination was a NESS cause of the adverse employment action. If the plaintiff meets its burden of production, then the claim is decided by the finder of fact—requiring that the plaintiff meet its burden of persuasion and prove beyond a preponderance of the evidence that the defendant's discrimination is a NESS cause of the adverse employment decision.[167]

Finally, in contrast with Title VII's burden-shifting scheme, a plaintiff's inability to prove but-for causation does not preclude her ability to recover damages.[168] Rather, proving NESS causation in the factorial framework, unlike proving that a protected characteristic is a motivating factor in Title VII's current burden-shifting scheme, permits the full range of damages that would be available to a plaintiff who proved but-for causation.[169]

[166] All evidence of causation under the potential-outcomes framework is circumstantial in the sense that it requires inferences regarding potential outcomes associated with counterfactual states of the world.

[167] Importantly, the proposed approach is not intended to affect the ability of a court to award injunctive relief to stop ongoing discriminatory conduct—a legislative issue separate from the current proposal. Instead, the focus of the proposed approach is on private actions for damages.

[168] *See supra* notes 37–39 and accompanying text.

[169] *See generally supra* Section II.A. Congress and the courts have been very reluctant to apply the motivating-factor test as a standalone, or even primary, standard of causation. Rather, in the narrow circumstances in which courts allow a plaintiff to rely on the motivating-factor test, the law severely limits the relief available to the plaintiff. *See supra* Section II.A; *see also* Babb v. Wilkie, 140 S. Ct. 1168, 1177–78 (2020) (limiting relief to "injunctive or other forward-looking relief" for "plaintiffs who demonstrate only that they were subjected to unequal consideration").

## B. The Centrality of the Potential-Outcomes Framework

The potential-outcomes framework plays a central role in both the theoretical and the practical components of the proposed approach to causation. It refines the causation inquiry by defining precisely what is meant by counterfactual causation, by providing a robust theoretical framework for asking and answering causal questions, and by creating a more seamless connection with *evidence* of causation.[170] As explained above, this theoretical framework also provides a strong foundation for the use of NESS as the causal estimand and a principal feature of the factorial approach.

In addition to its centrality to the meaning of causation in the factorial approach, the potential-outcomes framework is fundamental to the proof scheme described in the previous Section and, in particular, to determining whether the standard of causation has been satisfied.

### 1. Thinking Precisely About Proof

Identifying the primitive elements of the causal inquiry is valuable for thinking clearly about what a plaintiff or a defendant needs to prove. A plaintiff must identify, implicitly or explicitly, the units (e.g., the plaintiff, or perhaps all female employees of a company), the precise treatment variable (e.g., sex), including its relevant values (e.g., female and male) and its timing (e.g., the moment the defendant employer perceived the plaintiff's sex), and the outcome variable (e.g., promotion) and its relevant values (e.g., promoted or not promoted). In some cases, identifying these primitives will be relatively straightforward and in others they will require more thought and judgment.[171]

Either way, these elements serve as the building blocks of the causal inquiry. In defending against the plaintiff's discrimination claim, the defendant may challenge the primitives identified by the plaintiff. If the defendant alleges a legitimate purpose, it must similarly identify with precision the primitive elements of the causal inquiry involving the legitimate purpose.

Moreover, the potential-outcomes framework permits a precise conceptualization of the causal inquiry *for factfinding*. Regardless of how straightforward it is to identify the primitives associated with the parties' respective causal arguments, specifying potential outcomes and defining a causal effect with reference to them is an important component of making good inferences with respect to the causal inquiry. This is particularly so when the inquiry involves multiple factors. In these cases, precisely identifying causal primitives and potential outcomes facilitates a clear understanding of how causal factors interact and what inferences are required—in terms of well-

[170] *See generally Counterfactual Causation*, *supra* note 15 (applying the potential-outcomes framework to improve standards of causation in law).

[171] *See infra* notes 179–97 and accompanying text.

defined potential outcomes—to satisfy the causal standard. In short, it facilitates an accurate causal determination. For example, in the illustration above involving Monica,[172] identifying primitives and potential outcomes allows the factfinder to build, figuratively or literally, a matrix, such as that depicted in Figure 4, that provides a breakdown of the potential outcomes associated with each combination of factors and facilitates clear comparisons among them.

Figure 4. The factorial framework requires that a factfinder make inferences regarding potential outcomes, and complete, explicitly or implicitly, a matrix similar to the 2x2 matrix depicted

| | **Male** | **Female** |
|---|---|---|
| **Strongly Qualified** | ? | ? |
| **Weakly Qualified** | ? | ? |

### 2. Distinguishing Covariates from Intermediate Variables

In addition to facilitating precision in proving causation, the potential-outcomes framework is central to attaining accurate causal determinations by allowing litigants, courts, and factfinders to distinguish covariates from intermediate variables. As explained above, covariates are background variables that cannot be affected by the treatment variables.[173] An individual's height and pretreatment level of education are generally good examples of covariates. Intermediate outcome variables ("intermediate variables" or "intermediate outcomes") are variables that may be affected by the treatment but that are not the outcome variable of interest in the case.[174] Intermediate variables are important because a treatment may have a causal effect on an outcome variable via, or somehow associated with, its impact on an intermediate variable. For example, if an employee alleges that her employer intentionally discriminated against her based on her race by paying her less than her white colleagues, and she specifies race as the treatment variable and salary as the outcome variable, she may identify intermediate variables based on promotion or evaluation. It may be critical to distinguish these variables from covariates, such as a plaintiff's pretreatment level of education.[175]

[172] *See supra* Section IV.A.

[173] *See supra* Section III.C.2.

[174] *See supra* Section III.C.2.

[175] *See* Greiner, *supra* note 95, at 577–78.

Indeed, it is of central importance to distinguish covariates from intermediate variables for two reasons: (a) making appropriate comparisons, and (b) identifying preemptive causes.

### *i. Appropriate Comparisons*

When comparing one unit to other units for purposes of estimating the impact of a protected feature on an outcome variable (such as promotion or salary), it is appropriate to condition on relevant covariates. It is appropriate to compare individuals with and without the protected feature, conditional on them all having approximately the same relevant covariates. For example, a plaintiff may wish to show that an employer discriminates by tending to choose white candidates for hire over African-American candidates. The plaintiff's argument will be far stronger if, for example, his evidence involves African-American candidates having the same (or better) qualifications than the white candidates that were hired over them, and weaker if he cannot demonstrate that the African-American candidates had qualifications that were at least as strong as their white counterparts.

On the other hand, when making such comparisons—whether with statistical evidence involving a large dataset or just individual comparisons —it is not appropriate to condition on intermediate variables. A protected feature's effect on an outcome variable may, for example, operate via the intermediate variable, and conditioning on the intermediate variable may lead to a determination of no effect when one in fact exists.[176] For example, imagine a company that has a pay scale based on level of seniority in the company's hierarchical structure. A female employee starts her employment in the same position as a male employee. Five years later, the male employee is earning substantially more than the female employee because he has been promoted four times, whereas the female employee has only been promoted once. Conditioning on number of promotions (or level of seniority) would be a mistake. Promotion is an intermediate variable, and although the outcome variable is salary, conditioning on number of promotions likely ignores exactly the operation through which the employer discriminates against female employees and, as a consequence, pays them less. It would be invalid for the defendant employer to argue that the plaintiff must prove disparate salaries only while comparing female employees to male employees *of the same level of seniority*. Female employees in this scenario are paid less precisely *because* the employer discriminates against them on the basis of sex by failing to promote them at the same rate as their male counterparts.

The distinction between covariates and intermediate variables is crucial to any form of comparison—whether through individual comparison evidence or complex statistical analysis involving regression or other

[176] *See id.* at 576–80.

inferential methods. And, as Professor Greiner has stated, "classifying variables as covariates or intermediate outcomes . . . requires the analyst to understand the data-generating process thoroughly."[177] Indeed, it requires a framework, such as the potential-outcomes framework, for precisely defining the causal inquiry and its primitive elements.[178]

### *ii. Preemptive Causation*

The distinction between covariates and intermediate variables is also fundamental to conceptualizing the causal problem in the first instance, and, in particular, to determining whether a legitimate purpose is "preemptive" of an alleged discriminatory factor.[179] If a legitimate purpose is a covariate and is sufficient for the occurrence of the adverse employment action —presumably, it is alleged to be sufficient, since sufficiency is required to defeat but-for causation—then it is preemptive of the discriminatory factor. In other words, the outcome variable will have been determined prior to the unit's exposure to the discriminatory factor, implying that the discriminatory factor cannot be a cause of the outcome, the adverse employment action. As indicated above, our common intuition, the potential-outcomes framework, the NESS test, and other models of cause and effect, all tell us that a prerequisite of a treatment causing an outcome is the condition that exposure to the treatment must have preceded the outcome.[180]

The potential-outcomes framework permits a clear distinction between covariates and intermediate variables, and this distinction, in turn, allows an understanding of when the outcome variable is "fixed" (i.e., determined) prior to the unit's exposure to treatment—that is, of when the discriminatory factor is preempted by a legitimate purpose. In particular, a legitimate purpose is a preemptive cause of the adverse employment action when it is a sufficient covariate, a variable that cannot be affected by treatment and that is sufficient for the occurrence of the adverse outcome.[181]

This reasoning provides a good explanation for two elements of the factorial framework's prima facie case adopted from the *McDonnell Douglas* proof scheme—the requirement that a plaintiff establish the existence of an open employment opportunity controlled by the defendant and show that she met minimal qualifications for the opportunity.[182] Both of these elements

[177] *Id.* at 580.

[178] *See generally id.* at 576–80.

[179] *See supra* notes 156–60 and accompanying text; Wright, *supra* note 55, at 1794–98 (explaining preemptive causation).

[180] *See supra* notes 156–60 and accompanying text.

[181] To constitute a cause, the preemptive factor would also need to satisfy a necessity condition. For purposes of our discussion of preemption, however, what is important (regardless of whether the legitimate purpose satisfies such a condition) is whether the factor is a covariate and is sufficient for the adverse outcome.

[182] *See supra* Section IV.A.

reflect legitimate purposes that *preempt* the possibility of any discriminatory effect. They entail an outcome that precedes the treatment.

This component of the factorial framework is central to its sensibility. For example, imagine a racist gas station attendant who works as an *employee*, but who would not hire an African-American applicant if he were *hypothetically* the owner of the gas station. Imagine that he even admits to this openly. Applying a NESS standard of causation without concern for preemption, the gas station attendant may be found liable for discriminating against an African-American plaintiff by not hiring him. After all, there is good evidence that, had there been a job opening controlled by the defendant, the defendant would not have hired the plaintiff on account of his race. But a finding of liability in this scenario is illogical because there was no job opening controlled by the defendant in the first instance. The defendant did not even own the business or have other authority through which to offer a job.

Similarly, imagine a case in which the plaintiff alleges discrimination on the basis of sex, but in which the plaintiff, an applicant for a delivery-driver position, did not even have a driver's license or know how to drive. In this case, the legitimate purpose—not meeting minimal qualifications for the position—preempted any possible discriminatory effect. In other words, the outcome was entirely determined prior to the treatment. The job offer was foreclosed prior to the employer's perception of the plaintiff's sex.

Thus, when an alleged legitimate purpose preempts the discriminatory factor, there is no causation. Two main categories of preemptive causes are reflected in the *McDonnell Douglas* factors—the absence of an open employment opportunity and the absence of minimal qualifications. More broadly, however, this occurs when the legitimate purpose constitutes a sufficient covariate, in which case the outcome of the theoretical employment decision is determined prior to treatment. In other words, the employment opportunity is foreclosed prior to the defendant's perception of the protected feature.

But consider again the distinction between a covariate and an intermediate variable: a covariate is a variable that cannot be affected by treatment, whereas an intermediate variable can be affected by treatment. It is frequently nontrivial to determine whether a legitimate purpose is a covariate or an intermediate variable. For example, imagine that in response to allegations of discrimination based on sex, an employer acknowledges a job opening and the plaintiff's satisfaction of minimal qualifications, but asserts that the plaintiff's qualifications were worse than those of another applicant, who happened to be male. In effect, as is frequently the case, the employer alleges a legitimate purpose that involves a set of values representing different aspects of the plaintiff's qualifications (relative to those of her male counterpart) rather than a single value.

First, remember that a treatment is defined as an intervention *at a particular point in time.*[183] The time element—among the primitive components of the causal inquiry (implicit in the treatment and unit elements)—is fundamental to understanding whether a variable (in this case, a set of values) is a covariate or intermediate variable. In this example, we need to consider whether the treatment, the perception of the plaintiff's sex, came before or after the determination—or *instantiation*[184]—of the plaintiff's qualification values relative to those of her male counterpart. Are these objects "random variables," having multiple potential values, or are they rather fixed values by the time the treatment occurs?

At first glance, it may seem as though the plaintiff's qualifications are determined prior to her job application, which would imply that the set of values representing her qualifications would be a covariate. But this is not the case. It is the employer's *perception* of the plaintiff's qualifications relative to those of her male counterpart that is relevant to the inquiry.[185] The alleged legitimate purpose is based on this perception, and this perception may well be affected by the employer's perception of the plaintiff's sex. In other words, the defendant's assertion of a legitimate purpose based on the plaintiff's qualifications relative to those of her male counterpart likely involves an intermediate variable rather than a covariate.

To be sure, the question of whether a legitimate purpose constitutes a covariate or intermediate variable is a matter of judgment.[186] It depends on the particulars of the legitimate purpose. Whether it is a covariate or intermediate variable depends on whether it can be affected by the defendant's perception of the plaintiff's protected trait. And this is a matter of judgment.

As an illustration, distinguish the foregoing legitimate purpose—the plaintiff's weak qualifications relative to another applicant's qualifications —from a legitimate purpose reflected in the *McDonnell Douglas* factors, the failure of the plaintiff to *minimally* qualify for the employment opportunity. For example, let us return to the case of a delivery-driver applicant who does not have a driver's license. Like the weak-qualifications allegation, it must be determined whether this feature of the plaintiff, the absence of a driver's license, is subject to the post-treatment perception of the employer. Specifically, can the employer's perception of the plaintiff's sex affect his perception of whether the plaintiff meets the minimal qualifications of the open position? The answer: likely not.

[183] *See supra* note 96 and accompanying text.

[184] This term is borrowed from the broader causation literature. *See* Wright & Puppe, *supra* note 78, at 466–73.

[185] *See supra* notes 151–54 and accompanying text.

[186] Greiner, *supra* note 95, at 578–79; *see* Greiner & Rubin, *supra* note 95, at 775–78.

As another illustration, consider a scenario in which an elderly man asks a potential employer for a job and is rejected, allegedly on the basis of age. If it is determined that no open job existed at the time the elderly man requested a job—in the sense that the defendant's perception of the plaintiff's age could not have impacted whether a job opening became available—then the condition "no open position" is a sufficient covariate and preempts any possibility of discrimination. This may occur, for example, if, as above, the defendant did not own a business or have any other authority through which to offer a job. It may occur even if the defendant did own a business but was in no position to hire a new employee. This is a matter of judgment. On the other hand, if the defendant is an employer who has a history of impromptu hiring, then it is possible that the employer's perception of the plaintiff's age indeed affected his decision whether to offer the plaintiff a job. In this case, whether an employment opportunity existed may be better understood as an intermediate variable rather than a covariate, thus supporting an argument that a discriminatory factor was not preempted by the "no open position" condition.[187]

In short, fundamental to assessing whether a discriminatory factor was preempted by a legitimate purpose, and to determining causation more generally, is the ability to think sharply about the causal problem.[188] This is accomplished by carefully identifying the primitives of the causal inquiry and otherwise employing the potential-outcomes model to define the causal question and structure the ensuing analysis.

### 3. Developing Credible Statistical Evidence

Finally, the factorial framework's application of the potential-outcomes model permits better use of statistical data for determining causal effects. Indeed, the potential-outcomes model is a predominant framework for causal inference in statistics and the sciences.[189] Once a researcher has applied this model to define the causal inquiry and its primitive elements, she can apply the model to develop credible statistical evidence.

The primary concern of this Article is the *standard* of causation in disparate-treatment claims and not the assessment of whether certain data demonstrate satisfaction of this standard. It is therefore beyond the scope of this Article to explain the many benefits of applying the potential-outcomes framework to develop evidence of discrimination. Professor Greiner has provided a detailed explanation in this regard.[190] Suffice it to say, causal-

[187] Two main categories of preemptive legitimate purposes are those reflected in the *McDonnell Douglas* test and the factorial approach's prima facie case. However, others are possible also.

[188] *See* Greiner, *supra* note 95, at 576–80.

[189] *See generally* Section III.C.

[190] *See* Greiner, *supra* note 95.

inference methods currently employed in the courtroom are frequently outdated and not in line with modern techniques used in the sciences. More broadly, there are very good arguments for employing the potential-outcomes framework to prove or disprove discrimination statistically.[191] Indeed, it is not a stretch to say that current methods frequently lead to results that are simply not credible, and that the potential-outcomes framework may allow credible empirical analysis involving causal inference.[192] Among other things, the potential-outcomes framework provides a theory and method for deciding how to treat variables that are relevant to the causal inquiry (such as covariates), for understanding how to use data to make valuable comparisons, and for interpreting a study's results in common-sense terms.[193] It also facilitates neutral, unbiased—and, ultimately, more credible—analysis by providing a framework in which the researcher can more easily specify important decisions regarding the study's methodology prior to analyzing the data.[194] This prevents a wide range of questionable practices in which the researcher can manipulate her methodology in order to obtain results that favor her position.[195]

Moreover, even if the potential-outcomes framework is not adopted for purposes of analyzing data, employing the framework simply to structure the causal inquiry—for example, defining primitives of the causal inquiry and distinguishing covariates from intermediate variables—will benefit statistical analyses that are performed to prove or disprove causation. This is because, among the most damaging aspects of current methods of causal inference in the courts is the absence of a concrete, common-sense structure for defining suitable causal primitives and causal effects.[196] Instead, litigators and courts frequently rely blindly on regression analysis and "statistical minutiae," rather than important substantive decisions regarding foundational elements of a causal inquiry.[197]

### C. *The Role of Damages in Preventing Windfall Recoveries*

In the proposed approach, the damages phase of the litigation plays an important role in preventing windfall recoveries.

We are concerned here with windfall recoveries in a particular subset of disparate-treatment cases. When a plaintiff fails to make her case per the proof

[191] *Id.* at 535–39.

[192] *See* Greiner, *supra* note 95; s*ee also* Hillel J. Bavli, *Credibility in Empirical Legal Analysis* (SMU Dedman Sch. of L., Legal Studies Research Paper No. 434, 2020), https://ssrn.com/abstract=3434095 [https://perma.cc/R7WD-RCHU]; Daniel E. Ho & Donald B. Rubin, *Credible Causal Inference for Empirical Legal Studies*, 7 ANN. REV. L. & SOC. SCI. 17 (2011).

[193] Greiner, *supra* note 95, at 535–39.

[194] *Id.*; *see* Bavli, *supra* note 192, at 28–36.

[195] *See* Bavli, *supra* note 192, at 9–20, 28–36; Ho & Rubin, *supra* note 192, at 27–28.

[196] Greiner, *supra* note 95, at 537–38.

[197] *Id.*

structure described above, there is no issue of windfall recovery. This includes cases in which a plaintiff fails to prove that discrimination was a NESS cause of the adverse employment action. It also includes cases in which a defendant shows that a legitimate purpose *preempted* any discrimination—thus avoiding the most extreme category of windfall recovery. In particular, the factorial framework addresses directly, within its causal standard and as part of the liability question, the most concerning form of windfall recovery: windfall recovery that occurs when a legitimate purpose preempts the alleged discrimination, such as when there is no open employment opportunity or when the plaintiff fails to meet the minimal requirements for an opportunity. In these situations, the discriminatory factor does not meet the factorial framework's standard of causation. Finally, we are also not concerned here with windfall recoveries (at least no more than in other torts actions) when a defendant fails to establish a sufficient legitimate purpose, regardless of whether discrimination is shown to be a NESS cause. This is because our particular concern for windfall recoveries in the disparate-treatment context arises from the possibility of two or more sufficient causes, where at least one of which is legitimate.

What remains for consideration is the subset of cases that involve concurrent multiple sufficient causes. Our concern here pertains to cases in which the defendant establishes a sufficient legitimate purpose that is not preemptive of the discriminatory factor, as well as a sufficient discriminatory factor. A typical example involves a case in which both the plaintiff's race and the plaintiff's poor job performance are factors sufficient for the non-promotion of the plaintiff. Pursuant to the factorial approach, the plaintiff could prove his case, notwithstanding a concurrent sufficient legitimate factor, by showing that, in the absence of the legitimate factor (his poor job performance) his race would have made the difference between being promoted and not being promoted. Although we are interested in disincentivizing the employer's discriminatory behavior, liability in such a case gives rise to windfall-recovery concerns because the plaintiff's poor performance alone would have been sufficient to prevent him from receiving a promotion. Allowing him to recover could therefore have various distorting effects on the incentives of employees and employers.[198]

Although potential windfalls resulting from preemptive causation are better addressed in the liability phase as a matter of causation, the damages phase of the litigation provides an appropriate structure for preventing this latter category of windfall recoveries. Using damages to prevent such recoveries does not detract from antidiscrimination law's deterrence or

[198] *See generally* Babb v. Wilkie, 140 S. Ct. 1168, 1177–78 (2020) ("Remedies should not put a plaintiff in a more favorable position than he or she would have enjoyed absent discrimination.").

fairness objectives. To the contrary, it is well-aligned with them. Moreover, using damages calculations to prevent windfall recoveries is well-aligned with the role of damages in tort law generally.

In computing damages, there are various ways to account for a legitimate purpose to mitigate the harms associated with windfall recoveries.[199] An in-depth discussion of possible approaches is beyond the scope of this Article. My purpose here, however, is to show that good approaches are available. One possibility is to view legitimate and discriminatory purposes as two factors in a damages framework similar to a comparative-negligence scheme.[200] Applying such a framework, the jury could apportion damages according to some measure of responsibility, and a court could apply jury instructions that, for example, incorporate concerns for compensating the plaintiff and deterring discrimination, as well as concerns regarding windfall recoveries.

It is important to realize that requiring a logical causal framework in the liability stage of the litigation does not imply that the same framework must be applied in the damages stage. There is nothing that limits a damages computation to be determined only by concerns regarding causation. It is reasonable for it to be affected by a range of policy concerns.

On the other hand, there is logic in apportioning damages based on causation. In line with this logic, there are rigorous methods for making such apportionments using the potential-outcomes framework, even in msc situations. These methods can be based on unconditional, as well as conditional, effects. For example, a court may calculate damages based, at least in part, on the unconditional main effects of a protected feature, averaged over legitimate factors; or, similarly, based on the proportion of factor sets (or treatment combinations) involving the protected feature for which the protected feature is a necessary condition for the sufficiency of a set with respect to the adverse employment decision.

Assume, for example, that a female employee is paid less than her male colleague because of her sex and because of her job performance, each of which is a sufficient condition for her lower pay. Damages can be computed by considering four scenarios and associated potential outcomes: (1) those in

[199] One way of viewing windfall recoveries in discrimination cases is as follows: Allowing a plaintiff to obtain a windfall recovery is necessary in order to deter discrimination. Moreover, a *defendant* arguably obtains a windfall if he commits discrimination and is relieved from paying damages because the plaintiff happened to be less qualified or a poor performer. Using this model, a good damages calculation would properly balance these concerns. To the extent that the factorial framework could be criticized for employing damages calculations to mitigate the harmful effects associated with windfall recoveries, and balancing the foregoing concerns, rather than altogether avoiding these effects through liability determinations, this issue arises in all antidiscrimination proof schemes, since it reflects a tradeoff between windfalls for the plaintiff and windfalls for the defendant. Alternatively, it involves a tradeoff between avoiding windfalls and deterring discrimination.

[200] *See* Wright, *supra* note 55, at 1799 n.265; *see also* Clarke, *supra* note 148, at 783–85.

which the employee is not discriminated against but in which she has poor job performance; (2) those in which she is discriminated against but in which she does not have poor job performance; (3) those in which she neither is discriminated against nor has poor job performance; and (4) those in which she both is discriminated against and has poor job performance (i.e., the scenario that actually occurred).[201] For example, consider the potential outcomes in the 2x2 matrix in Figure 5.

Figure 5. 2x2 matrix depicting potential outcomes associated with combinations of two factors, sex and performance, each with two levels

| | **Male** | **Female** |
|---|---|---|
| **Strong Performance** | $32 | $24 |
| **Weak Performance** | $26 | $16 |

If the evidence leads the factfinder to arrive at these pay values, a damages calculation can be based on a main-effects analysis for the causal effect of "Female" versus "Male." This would involve a contrast between the values in the right column and the values in the left column of the 2x2 matrix. As a simple example, a court may apportion damages based on the average difference between the "Male" values and "Female" values: $29 – $20 = $9 (per hour). Contrast this $9/hour value with one based simply on a comparison between the employee's pay ($16) and that of her male colleague ($32), which would lead to a calculation based on the value $32 – $16 = $16.

A main-effects analysis can similarly apply for binary variables, such as promotion status. For example, if, instead of pay, the adverse employment decision was non-promotion, and the potential outcomes are found to consist of those in Figure 6, a damages calculation could similarly be based on a main-effects analysis that compares the proportion of factor sets, or treatment combinations, for which the adverse decision occurs when sex = Female to the proportion of treatment combinations for which the adverse decision occurs when sex = Male. Using the potential outcomes in Figure 6, the calculation would be based on the fact that promotion occurs in zero percent of the treatment combinations when sex = Female as compared to 50 percent when

[201] *See* Donald B. Rubin, *Estimating the Causal Effects of Smoking*, 20 STAT. MED. 1395, 1410–12 (2001) (examining the causal effects of alleged misconduct by the tobacco industry, and analyzing damages apportionment through consideration of "counterfactual worlds" when there are allegations of two "distinct sources of alleged misconduct"—misconduct by the tobacco industry and misconduct by the asbestos industry).

sex = Male.[202] In NESS terms, the calculation could be based on the proportion of factor sets with sex = Female for which sex = Female is a necessary condition for the sufficiency of a set with respect to the adverse employment decision "Not Promoted."

Figure 6. 2x2 matrix depicting (binary) potential outcomes associated with combinations of two factors, sex and performance, each with two levels

| | **Male** | **Female** |
|---|---|---|
| **Strong Performance** | Promoted | Not Promoted |
| **Weak Performance** | Not Promoted | Not Promoted |

Suffice it to say that in addition to preventing the most concerning form of windfall recovery by precluding causation when a legitimate purpose preempts a discriminatory factor, courts employing the factorial approach can use damages as a tool for further mitigating the harmful effects of windfall recoveries.

## V. Implications

The factorial framework carries a wide range of implications for disparate-treatment claims and antidiscrimination law generally. In this Part, however, I focus on two important implications in particular. First, in Section V.A, I build on the discussion above to argue explicitly that the factorial approach yields results that are more consistent with antidiscrimination law's policy objectives, and its deterrence objectives in particular, than either the motivating-factor test or the but-for test. Although the analysis herein involves antidiscrimination law's fairness (and compensation) aims also, I focus primarily on its deterrence objectives because the implications for these objectives flow less directly and are less apparent from the analysis elsewhere in this Article. Then, in Section V.B, I explain how, contrary to current methods, the proposed framework addresses the state of disarray surrounding causation in antidiscrimination law by simultaneously satisfying the causal language in antidiscrimination statutes and fulfilling Congress's broader aims in enacting these statutes—a task not possible using methods based on the motivating-factor or but-for tests.

### *A. The Deterrence Objectives of Antidiscrimination Law*

The factorial approach refines causation standards in discrimination cases. As discussed in Part II, these standards lack consistency and are

[202] *See supra* Section III.C; *Counterfactual Causation*, *supra* note 15, at 905–11.

inadequate in various important respects. The factorial framework, on the other hand, supplies a standard that fulfills the policy goals of antidiscrimination law, reflects dominant notions of actual cause and effect, and permits a logical and practical approach that can be applied consistently throughout the various areas of antidiscrimination law.

The factorial framework blends three innovations to accomplish a coherent and comprehensive standard. It employs: (1) the potential-outcomes model as a central structure in which to define and analyze the causal inquiry; (2) a main-effects analysis, and the NESS test in particular as its causal estimand; and (3) a legal framework grounded in tort law and recent innovations regarding multiple sufficient causes. In the previous parts of this Article, I have discussed benefits associated with each of these elements. In this Section, I discuss the implications of the factorial approach for antidiscrimination law's primary policy objective—deterring discrimination and, more broadly, encouraging socially desirable behavior.

The precise role of factual causation in fulfilling the law's policy objectives is complex. "A common belief . . . is that policy considerations have no role to play in the determination of cause-in-fact, 'because no policy can be strong enough to warrant the imposition of liability for loss to which the defendant's conduct has not *in fact* contributed.'"[203] The requirement of a causal link between misconduct and harm is at least firmly grounded in policy objectives of fairness and compensation. It is a well-accepted principle of tort law that individuals should only be held liable for harm that they have caused, and that individuals should only receive compensation grounded in tort for misconduct that has made some difference—that has been outcome determinative in some respect or other.[204] However, the role of causation in deterring harmful behavior, or more generally, in producing socially optimal behavior, has been more controversial. Law-and-economics scholars have long debated the role of factual causation in creating incentives for socially desirable behavior, or whether causation is even necessary for achieving such objectives in the first instance.[205]

In this Article, I take the position that antidiscrimination law seeks to accomplish a number of policy aims, including, most prominently, the goals of deterrence and fairness.[206] I assume that causation is of fundamental

[203] Viator, *supra* note 77, at 1526–27 (quoting JOHN G. FLEMING, THE LAW OF TORTS 170 (6th ed. 1983)).

[204] *See Counterfactual Causation*, *supra* note 15, at 892–93 nn.49–52 and accompanying text.

[205] *See generally* LANDES & POSNER, *supra* note 77; Landes & Posner, *supra* note 77; Shavell, *supra* note 77; Calabresi, *supra* note 77.

[206] *See* Price Waterhouse v. Hopkins, 490 U.S. 228, 264–65 (1989) (O'Connor, J., concurring) ("Like the common law of torts, the statutory employment 'tort' created by Title VII has two basic purposes. The first is to deter conduct which has been identified as contrary to public policy and harmful to society as a whole. As we have noted in the past, the award of backpay

importance to the policy objectives of tort law and antidiscrimination law in particular. This assumption is strongly supported by case law and scholarship, whether the importance of causation is rooted only in notions of fairness or also in deterrence theories.[207]

I therefore consider the implications of the factorial framework for the law's deterrence objectives, but with the important constraint that factual causation is a requirement of liability. In other words, I assume that, based *at least* on fairness and other non-deterrence objectives of antidiscrimination law —if not deterrence objectives also—a standard of causation that meaningfully reflects actual, common, and scientific cause and effect is necessary. The question is, what standard to apply?

Additionally, I consider a range of deterrence objectives that reflect, more broadly, aims of producing socially desirable behavior rather than simply deterring discrimination. Tort law seeks to deter misconduct by potential tortfeasors, but it also seeks to prevent overdeterrence and other socially undesirable "side effects," such as distorting the incentives of potential tort *victims* to take appropriate levels of precaution. I assume that this concern for the larger picture—for incentivizing socially desirable behavior, rather than simply maximizing deterrence of a particular form of misconduct—also applies to the special case of antidiscrimination law.

### 1. The Motivating-Factor Test

Once one accepts the premise that a standard of factual causation is necessary and that it should reflect actual cause and effect, the motivating-factor test cannot be accepted as a legitimate standard of causation. It effectively does away with the causation requirement.[208] Furthermore, even if it is assumed that the motivating-factor test reflects actual cause and effect, the test cannot be a satisfactory test of causation for two other related reasons.

First, the motivating-factor test is vague, and it leaves the causal determination to the impulses of jurors—whether in favor of a plaintiff or a

to a Title VII plaintiff provides 'the spur or catalyst which causes employers and unions to self-examine and to self-evaluate their employment practices and to endeavor to eliminate, so far as possible, the last vestiges' of discrimination in employment. The second goal of Title VII is 'to make persons whole for injuries suffered on account of unlawful employment discrimination.'" (citations omitted) (quoting Albemarle Paper Co. v. Moody, 422 U.S. 405, 417–18 (1975))).

[207] *See supra* Section II.A; David W. Robertson, *Causation in the* Restatement (Third) of Torts*: Three Arguable Mistakes*, 44 WAKE FOREST L. REV. 1007, 1008 (2009) ("[T]he cause-in-fact requirement is the 'linchpin' of the corrective-justice theory. Indeed, it has long been regarded as a truism that 'a defendant should never be held liable to a plaintiff for a loss where it appears that his wrong did not contribute to it, and no policy or moral consideration can be strong enough to warrant the imposition of liability in such [a] case.'" (second alteration in original) (first quoting Larry A. Alexander, *Causation and Corrective Justice: Does Tort Law Make Sense?*, 6 LAW & PHIL. 1, 12 (1987); then quoting Charles E. Carpenter, *Concurrent Causation*, 83 U. PA. L. REV. 941, 947 (1935))); *see also Counterfactual Causation*, *supra* note 15, at 884–85 n.14 and accompanying text.

[208] *See supra* Section III.A.

defendant. As such, it is likely to cause substantial uncertainty and a range of other undesirable effects that are avoidable with standards that provide more structure for the causal determination. The uncertainty that the motivating-factor test generates may have a range of unintended effects on employee and employer incentives. These include, for example, overdeterrence in the form of raising the costs of employment for potential employers—including insurance costs and social and financial risks for employers—and thereby causing fewer jobs and job-related opportunities. The uncertainty surrounding the standard of causation may also lead to suboptimal levels of litigation.

Second, overlapping with the issue of vagueness is the concern that the motivating-factor test is overbroad in the sense that it may allow a finding of causation even in circumstances in which an employer did not intend to discriminate. This overbreadth may also give rise to socially undesirable behavior. In particular, it may cause a range of unintended and detrimental effects on the incentives of employees and employers. For example, if an employer has been accused of committing discrimination in the past, he may avoid creating new jobs, giving promotions, or selecting employees (or perhaps even selecting *certain* employees) for opportunities in order to avoid potential exposure to liability based on his employment decisions, *even if he has no intention of committing discriminatory behavior*. This overdeterrence may occur because, under the motivating-factor test, evidence of his earlier discriminatory acts may be used, and may be sufficient, to show discrimination in his employment decision, even if he is able to prove that the applicants' qualifications were entirely determinative of his employment decision.[209]

Of course, a certain level of fear of legal exposure is ideal. This is how deterrence works. The law intends to threaten legal exposure for employers who discriminate. However, too much fear—such that the threat of liability deters not only discriminatory conduct, but also non-discriminatory, socially beneficial conduct—is counterproductive. In the extreme case, overdeterrence can eliminate jobs altogether, even when an employer has no intention to discriminate.

Finally, for both the vagueness and overbreadth issues discussed above, it is important to realize that overdeterrence constitutes only one category of the harms produced by the motivating-factor test. Another category pertains to *underenforcement* and *underdeterrence*. The motivating-factor test provides juries with little guidance and allows them great flexibility to arrive at decisions they find appropriate. In some cases, a jury may find an employer to have acted appropriately where, in fact, the employer acted discriminatorily.

[209] *See generally* Hillel J. Bavli, *An Aggregation Theory of Character Evidence* (SMU Dedman Sch. of L., Legal Studies Research Paper No. 483, 2020), https://ssrn.com/abstract=3664837 [https://perma.cc/Q4RQ-NK7G] (discussing other-acts character evidence, including evidence of prior discriminatory acts).

A jury in one part of the country may be far more likely to view certain conduct as discriminatory than a jury in another part of the country.

Moreover, an employer may be *underdeterred* if he believes that a jury is unlikely to hold him responsible for his discriminatory conduct. Because the motivating-factor test leaves the decision to the intuition of the jury, juries may rule with reference to their personal norms and intuitions, rather than the law's aims. It may be difficult for courts to control this tendency or overturn verdicts when the motivating-factor test sanctions this flexibility.

Thus, it is likely that the motivating-factor test both overdeters and underdeters, and that it creates poor incentives for both employers and employees, due to its inherent vagueness and its focus on forces that may "motivate" employer conduct rather than on the notion of "making a difference" employed in well-established conceptions of cause and effect.

### 2. But-For Causation

On the other hand, the but-for test also fails to accomplish optimal deterrence. As discussed, it has many advantages: it is simple and straightforward, it employs an explicit analytical process, and it reflects actual cause and effect. Its major weakness, however, is that it underdeters discriminatory conduct by allowing employers to evade liability through arguments that there were sufficient legitimate causes.

Consider, for example, a scenario in which an employer was hiring a new employee for his business. Two individuals submitted applications for the position—a white applicant, who was incidentally very well-qualified for the position, and an African-American applicant, who was incidentally not well-qualified for the position. Upon interviewing the African-American applicant, however, the employer made no reference to the applicants' qualifications. Instead, the employer made it clear that he simply would not hire a non-white applicant.

The antidiscrimination laws undoubtedly aim to deter such behavior. However, it is unlikely that the employer's conduct would give rise to liability under the but-for test because in the absence of the employer's discriminatory conduct, the employer is very unlikely to have hired the African-American applicant anyway. Knowing this, employers are likely to be more willing to engage in various forms of discriminatory behavior.

### 3. The Factorial Framework

The factorial framework addresses the deterrence problems discussed above in a straightforward way. It employs the NESS estimand, which asks whether the discriminatory conduct was necessary for the sufficiency of a set of factors for the occurrence of the adverse employment decision. Using the formulation in the Restatement (Third) of Torts, the NESS estimand applies to the foregoing scenario to require the following finding of fact: had the candidates been similarly qualified, would the plaintiff's race have made a difference in the employer's employment decision—would the plaintiff being African American versus white be likely to have made a difference? The answer to this question, in light of the facts above, is clearly "yes." Therefore, applying the factorial framework, the employer's discriminatory conduct is a cause of the adverse employment action.

Note that implicit in this analysis are numerous decisions that would require careful thought using the potential-outcomes framework—for example, whether the treatment levels associated with the factor "race" consist of "white" and "non-white" or "white" and "African American," and whether the plaintiff's poor qualifications constitute sufficient covariates and preempted the employer's discrimination by failing to meet minimal requirements.[210]

As the example above demonstrates, the NESS test leads to liability in certain situations in which the but-for test underdeters. It addresses the but-for test's underenforcement problem. Let us now consider how the NESS test would operate in an example referred to above to demonstrate the motivating-factor test's overenforcement and overdeterrence problems.[211] Assume that an employer has been accused of discriminatory behavior in the past and, although he has no intention to discriminate in seeking a new employee, he is hesitant to establish a new position because he fears that allegations of his past discriminatory conduct may be used to prove that race permeated his employment decision. He fears that, under the motivating-factor test, he may be found liable for discrimination even if he avoids any and all discriminatory behavior. This is problematic for the reasons discussed above.[212] However, unlike the motivating-factor test, the NESS test avoids this concern, or at least accounts for it while balancing concerns regarding an overly stringent causation standard.

Again using the Restatement (Third)'s formulation, the NESS test would ask whether, absent differences in qualifications, race would have made a difference. Unlike the motivating-factor test, which may, based on allegations regarding the employer's past behavior, identify race as a factor that "motivated" his employment decision, the NESS test would require evidence that race would have *made a difference.* Without more, the allegations regarding the employer's past discriminatory behavior would likely be insufficient for a finding of causation using the NESS test. The factorial approach thus avoids certain overenforcement and overdeterrence concerns associated with this type of situation. Employing the potential-outcomes framework and the NESS test as its estimand, the proposed approach, while not requiring a different actual outcome, requires a difference in *potential* outcomes—in what would have happened.

[210] *See supra* Section IV.B.

[211] *See supra* Section V.A.1.

[212] *See supra* Section V.A.1.

In short, the factorial approach satisfies the criterion that a standard of causation follow the notion of actual cause and effect. At the same time, it does not suffer from the underenforcement and underdeterrence weaknesses of the but-for test. It also addresses the vagueness concerns and the substantial enforcement and deterrence problems associated with the motivating-factor test by providing a concrete measure of causation based on the counterfactual model and the necessity condition.[213] Finally, through careful design of the causal inquiry using the potential-outcomes framework, the factorial approach avoids overdeterrence and other socially undesirable effects that result from misclassifying a factor as a cause of an adverse employment action when such a factor is altogether preempted by a legitimate cause.

### 4. Effect Categories

Based on the above argument, it seems beneficial to rule out the motivating-factor and but-for tests in favor of the factorial framework. However, let us also consider a categorization of effects under the motivating-factor, but-for, and NESS tests to better understand the relationship between these tests and the benefits of the factorial framework.

Consider two categories of cases for which the results of the motivating-factor test and the but-for test differ from the results of the factorial approach. First, consider cases that exist within the nebulous area between the fuzzy threshold for satisfying the motivating-factor test and the threshold for a minimum level of "difference" that a protected feature can make—i.e., NESS. These are cases that satisfy the motivating-factor test but not the NESS test or the but-for test. These cases involve circumstances in which the protected feature did not even make a hypothetical difference, let alone an actual difference. They include circumstances in which a protected feature played

[213] Note that the proposed approach to causation may also have a wide range of implications for other areas of antidiscrimination law. For example, it may have important implications for disparate-treatment *class actions*. By providing a theory of counterfactual causation that is less stringent than the but-for test, on the one hand, and, in a sense, less individualistic than the motivating-factor test, on the other hand, the factorial framework may allow class treatment in contexts in which such treatment is currently foreclosed. It may also facilitate class treatment by enabling certain statistical methods, grounded in the potential-outcomes framework, for developing evidence of a "pattern or practice" of discrimination. *See generally* LINDEMANN ET AL., *supra* note 8, at 2-114–19. Indeed, it may have broader implications for proving pattern-or-practice cases—where a plaintiff seeks to show "a pattern or practice of disparate treatment"—and for burden-shifting schemes currently used in those cases. *Id.* at 2-115. Finally, the factorial framework may have implications for disparate *impact* cases, in which "an employer's facially neutral policy or practice may be unlawful—even absent a showing of discriminatory intent—if it has a significant disparate impact on a protected group." *Id.* at 3-2 (citing Griggs v. Duke Power Co., 401 U.S. 424, 430–32 (1971)). In these cases, once a plaintiff has established a prima facie case involving disparate consequences of an employer's policy or practice, "the employer may defend its policy or practice by proving that it is 'job related for the position in question and consistent with business necessity.'" *Id.* at 3-2, 3-37 (quoting 42 U.S.C. § 2000e-2(k)(1)(A)(i) (2018)).

such an insignificant role, if any at all, that, even in the absence of the other factors (e.g., even if two applicants had been equally qualified), the protected feature would not have made a difference in the outcome of the employment decision.

For the reasons discussed above, findings of causation for this category of cases are not optimal. Even if there is some deterrence benefit associated with findings of causation in these cases, there is little question that the harmful effects associated with the ambiguity surrounding these findings under the motivating-factor test, combined with the harmful effects of overdeterrence, would outweigh any theoretical benefit.

Second, consider a category of cases that do not meet the but-for standard, but that do meet the NESS standard. In particular, consider a subcategory of these cases for which the motivating-factor test provides a jury with flexibility to find no discrimination even when discrimination constitutes a NESS cause. Findings of no liability in these cases, however, seem out of line with the intention of the motivating-factor test based on the test's own premise. After all, it is a form of the substantial-factor test in tort law, a test developed to find causation in msc situations when the but-for test could not. It is only through the vague nature of the motivating-factor test that a jury could render a finding of no causation under this test even when the alleged discrimination is a NESS cause of an adverse employment action. It is at least reasonable, if not necessary, to conclude that, even by the motivating-factor test's own measures and objectives, the factorial approach is preferable in these situations.

Let us continue to consider the category of cases that satisfy the NESS test but not the but-for test, but now with reference to the but-for test rather than the motivating-factor test. First, it seems almost obvious, or at least well-accepted, that enabling findings of liability in mixed-motive cases would serve the law's deterrence objectives. This conclusion is based on the near-universal rejection of the but-for standard in msc situations in tort law—a field of law that is particularly concerned with deterrence and incentivizing socially optimal behavior—and the substantial concern, articulated repeatedly by Congress, courts, and scholars, for deterring discriminatory behavior, even when that behavior is accompanied by sufficient legitimate purposes.[214]

As Justice O'Connor emphasized in her concurring opinion in *Price Waterhouse*, "There is no doubt that Congress considered reliance on gender or race in making employment decisions an evil in itself."[215] Moreover, the "Court's decisions under the Equal Protection Clause have long recognized

[214] *See supra* Part II.

[215] *See* Price Waterhouse v. Hopkins, 490 U.S. 228, 264–65 (1989) (O'Connor, J., concurring) (opining further that "[w]hile the main concern of [Title VII] was with employment opportunity, Congress was certainly not blind to the stigmatic harm which comes from being evaluated by a process which treats one as an inferior by reason of one's race or sex").

that whatever the final outcome of a decisional process, the inclusion of race or sex as a consideration within it harms both society and the individual."[216] Therefore, "[w]here an individual disparate treatment plaintiff has shown by a preponderance of the evidence that an illegitimate criterion was a *substantial factor* in an adverse employment decision, the deterrent purpose of the statute has clearly been triggered."[217]

Thus, as discussed above, and confirmed by Justice O'Connor's opinion, the but-for test underdeters, and the NESS test is preferable in this respect. The next question to ask, however, is whether the NESS test overdeters. Remember, we are not examining the general dangers of overdeterrence in disparate-treatment cases; rather, we are interested particularly in the overdeterrence that results from allowing liability in cases in which discrimination, although not a but-for cause, is a cause under the NESS test. This may occur, for example, where, in the absence of a sufficient legitimate factor, being male rather than female would have made the difference between being hired and not hired. But this category of cases consists of causes—multiple sufficient causes—that are decidedly underdeterred.[218] This much is clear from Supreme Court decisions and a long tradition of exceptions to but-for causation for cases involving multiple sufficient causes.[219] On the other hand, the NESS test is substantially less stringent than the but-for test, and with its added flexibility comes the potential for overdeterrence.

The NESS estimand has the potential to create suboptimal incentives by allowing windfall recoveries when, for example, an employee fails to meet minimal requirements for an employment opportunity, or when an adverse employment decision is a result of a plaintiff acting inappropriately or unproductively at work—although also a result of discrimination. As discussed above, the factorial framework addresses the potential for suboptimal incentives associated with this subset of cases in two ways.[220] First, the factorial framework precludes a finding of causation when the adverse employment decision is the result of a sufficient covariate—that is, when the discriminatory force is preempted by another force. Second, windfall recoveries that result from applying the NESS test to discrimination that is accompanied by a *nonpreemptive* sufficient legitimate factor can be mitigated (or optimized[221])

[216] *Id.* at 265 (citing Richmond v. J.A. Croson Co., 488 U.S. 469 (1989)). Interestingly, Justice O'Connor's opinion continues by highlighting that "Congress clearly conditioned legal liability on a determination that the consideration of an illegitimate factor *caused* a tangible employment injury of some kind." *Id.*

[217] *Id.*

[218] *See supra* notes 214–17 and accompanying text.

[219] *See supra* Part II.

[220] *See supra* Sections IV.B–.C.

[221] *See supra* note 199.

with an appropriate damages calculation. To be sure, employing damages calculations as a method for mitigating windfall recoveries is not perfect. For example, it lacks precision. However, it is flexible, and it is arguably the best of imperfect alternatives. Moreover, there is strong precedent for such calculations from the broader torts context.[222]

### *B. Causal Language in Antidiscrimination Statutes*

Antidiscrimination statutes generally contain causal language that courts have held to require a but-for standard. For example, the Supreme Court held in *Gross v. FBL Financial Services, Inc.* that the ordinary meaning of the ADEA's language, that an employer is liable for discrimination against an individual "because of such individual's age," is "by reason of" or "on account of" age —i.e., "that age was the 'but-for' cause of the employer's adverse decision."[223] The Court distinguished *Price Waterhouse* and rejected the argument that the Court should apply Title VII's burden-shifting proof scheme, highlighting "textual differences between Title VII and the ADEA" and Congress's choice to incorporate the burden-shifting scheme in the former statute while not in the latter.[224]

In *University of Texas Southwestern Medical Center v. Nassar*, the Supreme Court similarly held that Title VII's antiretaliation provision's use of the term "because" implies that "Title VII retaliation claims require proof that the desire to retaliate was the but-for cause of the challenged employment action."[225]

The causal language in these and other antidiscrimination statutes have caused immense confusion.[226] As the cases above suggest, there have been multiple Supreme Court decisions regarding whether such language in antidiscrimination statutes requires a but-for standard or permits a motivating-factor standard. Indeed, the Supreme Court recently handed down two

[222] *See supra* notes 200–01 and accompanying text.

[223] Gross v. FBL Fin. Servs., Inc., 557 U.S. 167, 176 (2009) (quoting 1 WEBSTER'S THIRD NEW INTERNATIONAL DICTIONARY 194 (1966)).

[224] *Id.* at 173–75 n.2; Univ. of Tex. Sw. Med. Ctr. v. Nassar, 570 U.S. 338, 350–51 (2013) (internal quotation marks omitted). Title VII also involves a but-for standard, but only as part of a more complex proof scheme. *See supra* Section II.A.

[225] *Nassar*, 570 U.S. at 352.

[226] Consider, for example, the causal language of the ADA. As with other statutes, courts vary significantly as to whether the ADA's "on the basis of" language implies a but-for standard or permits a motivating-factor standard. *See* Parker v. Sony Pictures Ent., Inc., 260 F.3d 100, 105 –08 (2d Cir. 2001); Serwatka v. Rockwell Automation Inc., 591 F.3d 957, 961–64 (7th Cir. 2010); Gulliford v. Schilli Transp. Servs., Inc., No 4:15-CV-19-PRC, 2017 WL 1547301, at *6 (N.D. Ind. Apr. 27, 2017); *see also Counterfactual Causation*, *supra* note 15, at 930 n.152 and accompanying text; LINDEMANN ET AL., *supra* note 8, at 13-197–99.

decisions regarding the meaning of causal language in antidiscrimination statutes and its implications for causation.[227]

The reason for this turmoil is simple: the causal language in these statutes is thought to imply but-for causation, whereas good policy, common sense, and the intent of Congress is thought to require a less stringent standard. In other words, there is substantial discordance between the statutes' causal language, as it is currently understood, and the broader intent of Congress.

The factorial approach addresses this discordance. Through the potential-outcomes framework and the NESS estimand, it provides a causal measure that is consistent with both the causal language of the antidiscrimination statutes on the one hand, and good policy and the intent of Congress on the other.[228] This harmonization is reflected in this Article's characterization of the factorial framework as a refinement of both the but-for test and the motivating-factor test.

For example, a central issue in cases recently decided by the Supreme Court is whether, pursuant to causal statutory language, the alleged discriminatory conduct must have "made a difference" in order to satisfy causation.[229] In these decisions, the Court presumes that requiring this necessity condition implies a but-for standard and precludes the motivating-factor test. However, the factorial framework rejects this presumption. A central feature of the proposed approach is the notion that a force may make a difference—it may fulfill the necessity condition—without going so far as to satisfy the but-for standard. Pursuant to this approach, a logical and effective standard of causation may indeed require the necessity condition, but in a less stringent form than the but-for test.

My argument here relies on two components: (1) the ordinary meaning of the causal language in antidiscrimination statutes reflects the ordinary meaning of cause and effect, which, in turn, is intended to capture the scientific meaning of cause and effect, including both conditional effects and unconditional effects, and the NESS estimand in particular; and (2) the factorial framework is consistent with good policy and Congress's intent.

First, as explained in *Counterfactual Causation*, the ordinary meaning of causal language such as "because of" and "on the basis of" is easily interpreted as capturing the broader meaning of causation that includes both conditional effects and unconditional effects.[230] The Court "begin[s] with the language employed by Congress and the assumption that the ordinary meaning of that

[227] *See* Babb v. Wilkie, 140 S. Ct. 1168 (2020); Comcast Corp. v. Nat'l Ass'n of Afr. Am.-Owned Media, 140 S. Ct. 1009 (2020); *see also* Bostock v. Clayton County, 140 S. Ct. 1731 (2020); *infra* Section V.C.

[228] *See Counterfactual Causation*, *supra* note 15, at 925–32.

[229] *See Babb*, 140 S. Ct. at 1172; *Comcast Corp.*, 140 S. Ct. at 1013–14.

[230] *Counterfactual Causation*, *supra* note 15, at 902.

language accurately expresses the legislative purpose,"[231] and to determine the ordinary meaning of terms in a statute, the Court examines their entries in well-established dictionaries.[232] Further, it is clear from the Supreme Court's analysis of causal statutory language such as "because of," "on the basis of," and "results from" that the Court's interpretation of those terms is based on the common meaning of *cause and effect.* For example, in *Burrage v. United States*, a criminal case involving the Controlled Substances Act, the Court interpreted the ordinary meaning of the Act's "results from" language, stating:

> A thing "results" when it "[a]rise[s] as an effect, issue, or outcome from some action, process or design." 2 The New Shorter Oxford English Dictionary 2570 (1993). "Results from" imposes, in other words, a requirement of actual causality. "In the usual course," this requires proof "'that the harm would not have occurred' in the absence of—that is, but for—the defendant's conduct." University of Tex. Southwestern Medical Center v. Nassar, 570 U.S. ——, ——, 133 S. Ct. 2517, 2525, 186 L.Ed.2d 503 (2013) (quoting Restatement of Torts § 431, Comment a (1934)).[233]

Similarly, in *Gross*, the Court held that "because of . . . age" means "that age was the 'reason' that the employer decided to act."[234] Terms such as "because of," "on the basis of," and "results from" imply a requirement of actual cause and effect, which, in turn, courts interpret to mean but-for causation.[235]

But, let us now consider whether actual cause and effect in fact implies but-for causation. Examining dictionary entries for the term "causation" reveals that the ordinary meaning of the term is intended to capture, or is at least extremely well correlated with, the scientific concept of cause and effect, which includes both conditional and unconditional effects.[236] For example, Merriam-Webster defines "causation" as "the act or process of causing"; "the act or agency which produces an effect."[237] Oxford Dictionaries defines the term as "[t]he action of causing something"; "[t]he relationship between

[231] Engine Mfrs. Ass'n v. S. Coast Air Quality Mgmt. Dist., 541 U.S. 246, 252 (2004) (quoting Park 'N Fly, Inc. v. Dollar Park & Fly, Inc., 469 U.S. 189, 194 (1985)).

[232] *Counterfactual Causation*, *supra* note 15, at 902 (citing Burrage v. United States, 571 U.S. 204, 210–11 (2014); Gross v. FBL Fin. Servs., Inc., 557 U.S. 167, 176 (2009)).

[233] *Burrage*, 571 U.S. at 210–11.

[234] *Gross*, 557 U.S. at 176.

[235] *See Burrage*, 571 U.S. at 211–12 ("[C]ourts regularly read phrases like 'results from' to require but-for causality.").

[236] *Counterfactual Causation*, *supra* note 15, at 903.

[237] *Causation*, MERRIAM-WEBSTER, https://www.merriam-webster.com/dictionary/causation [https://perma.cc/77EJ-5WKX].

cause and effect; causality."[238] Most significantly, the primary usage examples of this term provided by Merriam-Webster are "the role of heredity in the *causation* of cancer," and "in a complex situation *causation* is likely to be multiple"; and those provided by Oxford Dictionaries are "investigating the role of nitrate in the causation of cancer," and "a strong association is not a proof of causation."[239]

It is clear from these entries in well-established dictionaries—regularly referred to by the Supreme Court and lower courts to determine the ordinary meaning of causal and other statutory language—that the ordinary meaning of the causal language in antidiscrimination statutes incorporates the scientific meaning of cause and effect, which is broader than but-for causation. It includes *both* conditional effects (and but-for causation in particular) and unconditional effects, including the NESS estimand.[240]

Importantly, as emphasized above, neither the ordinary meaning nor the scientific meaning of causation includes the motivating-factor test.[241] However, the ordinary and scientific meaning of causation is broader than the but-for test: it includes the broader meaning of "making a difference" —of the necessity condition—reflected in a main-effects analysis.[242]

This is not to say that courts have necessarily been incorrect to interpret the ordinary meaning of causal statutory language to mean but-for causation. In this Article, I define but-for causation consistently with its common understanding in law—as a specific conditional effect within the broader counterfactual model.[243] However, the key component of the but-for test is the necessity condition, the condition that, in one form or another, the factor made a difference. This component reflects a comparison of potential outcomes. It is central to both conditional effects and unconditional effects. Therefore, when the courts apply but-for causation to reflect the ordinary meaning of cause and effect, they may simply intend to apply the concept of the necessity condition, which, in turn, applies to unconditional effects as well as conditional effects.

Similarly, when Congress uses causal language in antidiscrimination statutes, it is likely referring to the broader necessity condition rather than the narrow conditional effect I (and courts) have called but-for causation. Whether Congress was aware of this broader meaning or not, it is likely to have intended it. The broader meaning within the counterfactual framework

[238] *Causation*, LEXICO, https://www.lexico.com/en/definition/causation [https://perma.cc/QFE8-6GGX]; *see Counterfactual Causation*, *supra* note 15, at 903.

[239] MERRIAM-WEBSTER, *supra* note 237; LEXICO, *supra* note 238; *see Counterfactual Causation*, *supra* note 15, at 903.

[240] *See Counterfactual Causation*, *supra* note 15, at 903, 927–32.

[241] *See supra* Sections II.B, III.A.

[242] *See supra* Sections III.C–.E.

[243] *See supra* Part III.

is in fact captured by the "ordinary" meaning of cause and effect. Moreover, as suggested by courts and scholars—at least implicitly—Congress's objectives in enacting the antidiscrimination laws are better served by a causal standard reflecting a broader meaning of the necessity condition.[244]

Indeed, the discordance between the Supreme Court's interpretation of Congress's causal language as implying the but-for standard, on the one hand, and good policy and Congress's objectives in enacting the antidiscrimination statutes, on the other hand, is likely the primary reason for the state of disarray surrounding this area of the law. The theory and methodology provided by the factorial framework resolves this discordance. As explained in Part III, the broader meaning of counterfactual causation reflected in a main-effects analysis, and the NESS estimand in particular, comports with good policy and Congress's likely intent.

### C. *An Application: The Supreme Court's Decisions in* Comcast *and* Babb

In 2020, the Supreme Court handed down two decisions regarding the appropriate standard of causation in disparate-treatment cases.[245] In *Comcast Corporation v. National Association of African American-Owned Media*, Entertainment Studios Network (ESN), an operator of television networks, and the National Association of African American-Owned Media sued Comcast under 42 U.S.C. § 1981, alleging discrimination based on race in "mak[ing] or enforc[ing] contracts."[246] The Supreme Court, however, unanimously ruled that a plaintiff bringing an action for discrimination under 42 U.S.C. § 1981 must prove that race was a but-for cause of the adverse decision, and that the plaintiffs did not satisfy this standard.[247]

The Court reasoned that this "ancient and simple 'but for' common law causation test . . . supplies the 'default' or 'background' rule against which Congress is normally presumed to have legislated," and that "[n]ormally . . . the essential elements of a claim remain constant through the life of a lawsuit."[248] The Court held that, based on "the statute's text, its history, and [the Court's] precedent . . . § 1981 follows the general rule."[249] The Court used similar reasoning as in earlier cases, drawing on the meaning of terms

[244] *See supra* Section V.A; *see also supra* Sections II.A–.B.

[245] *See* Comcast Corp. v. Nat'l Ass'n of Afr. Am.-Owned Media, 140 S. Ct. 1009 (2020); Babb v. Wilkie, 140 S. Ct. 1168 (2020); *see also* Bostock v. Clayton County, 140 S. Ct. 1731 (2020).

[246] 42 U.S.C. § 1981 (2018); *Comcast Corp.*, 140 S. Ct. at 1013.

[247] *Comcast Corp.*, 140 S. Ct. at 1019.

[248] *Id.* at 1014.

[249] *Id.*

like "on account of," "by reason of," "on the basis of," and "because of," as well as the default causation standard in tort law, to infer a but-for standard.[250]

In *Babb v. Wilkie*, the Supreme Court considered a claim brought under 29 U.S.C. § 633a(a), the federal-sector provision of the ADEA.[251] The claim was brought (among other discrimination claims not considered by the Court) by Noris Babb, a clinical pharmacist at the Department of Veterans Affairs Medical Center ("VA"), against her employer, the VA.[252] Babb alleged that, as a result of age discrimination, the VA took away a designation that made her eligible for promotion, denied her certain training opportunities, and reduced her holiday pay.[253] She also alleged a number of age-related comments made by her supervisors.[254] In response, the VA alleged legitimate reasons for the adverse employment actions.[255]

Unlike the Court's holding in *Comcast*, the Court held that, here, "age need not be a but-for cause of an employment decision in order for there to be a violation of [the statute]."[256] The decision turned on the plain meaning of the statute's text. The relevant portion of the ADEA provides: "All personnel actions affecting employees or applicants for employment who are at least 40 years of age . . . shall be made free from any discrimination based on age."[257]

The Court pieced apart the statute's language, interpreting each term based on its plain dictionary meaning.[258] It then analyzed the relationships between the terms, and concluded that "age must be a but-for cause of discrimination—that is, of differential treatment—but not necessarily a but-for cause of a personnel action itself."[259] According to the Court, the "straightforward meaning" of the statute "does not require proof that an employment decision would have turned out differently if age had not been taken into account."[260] Rather, "[t]he plain meaning of the critical statutory

[250] *See id.* at 1014–16. The Court also distinguished § 1981 from Title VII, and highlighted that

> it's not as if Congress forgot about § 1981 when it adopted the Civil Rights Act of 1991. At the same time that it added the motivating factor test to Title VII, Congress *also* amended § 1981. But nowhere in its amendments to § 1981 did Congress so much as whisper about motivating factors.

*Id.* at 1017–18 (citation omitted).

[251] Babb v. Wilkie, 140 S. Ct. 1168, 1171 (2020).

[252] *Id.*

[253] *Id.*

[254] *Id.*

[255] *Id.* at 1171–72.

[256] *Id.* at 1172.

[257] 29 U.S.C. § 633a(a) (2018); *Babb*, 140 S. Ct. at 1172.

[258] *Babb*, 140 S. Ct. at 1173–74.

[259] *Id.* at 1173.

[260] *Id.* at 1174.

language ('made free from any discrimination based on age') demands that personnel actions be untainted by any consideration of age."[261]

The Court illustrated its reasoning with an example involving an employer that uses a point system to make promotion decisions: If a 55-year-old had a score of 85, and then a 5-point deduction based on age, whereas a 35-year-old had a score of 90, and therefore would have received the promotion regardless of the age-based point deduction, then the decision is not made "free from any discrimination," even though the point deduction ultimately did not change the outcome of the promotion decision.[262]

The Supreme Court concluded that a plaintiff who proves "unequal consideration," but is unable to show but-for causation, may "seek injunctive or other forward-looking relief," but "cannot obtain reinstatement, backpay, compensatory damages, or other forms of relief related to the end result of an employment decision."[263]

Both decisions are based primarily on the Court's interpretation of the plain meaning of the statutes on which the respective claims are based. As in earlier cases, the Court reads causal language such as "because of" and "based on" to imply a but-for requirement.[264] As discussed, this standard leads to problems in disparate-treatment cases. However, as explained in the previous Section, these terms likely entail a broader meaning—one more in line with NESS than but-for causation. Moreover, this standard would not lead to the problems associated with the but-for test, and therefore would not lead to dissonance between the literal meaning of the statute's language on the one hand, and good policy and Congress's likely intent on the other. For these reasons, the factorial framework would likely provide a more appropriate standard of causation in *Comcast* than the but-for test.

Moreover, respectfully, the Court's analysis in *Babb* is problematic in a number of respects. The factorial framework provides a theory and standard that resolves these issues. First, the Court's "any-consideration" standard[265] does not jibe easily with modern conceptions of cause and effect. It is a form of the motivating-factor test and is subject to all of the same criticisms that apply to that test generally. The any-consideration test that the Court illustrates in its example is a standard of *behavior*, not one of causation. Causation is the element of a legal claim that links misconduct with harm.[266] An employer who considers age in making an employment decision—such as the 5-point deduction in the Supreme Court's example—has committed an

[261] *Id.* at 1171 (quoting 29 U.S.C. § 633a(a)).

[262] *Id.* at 1174.

[263] *Id.* at 1177–78.

[264] *See, e.g.*, *supra* notes 248–50 and accompanying text.

[265] *See Babb*, 140 S. Ct. at 1179 (Thomas, J., dissenting).

[266] *See supra* notes 2–3 and accompanying text.

impermissible, discriminatory act. This is separate from the causation question, which asks whether that act is properly linked to the plaintiff's harm.[267] *This*, and not just but-for causation, is arguably "the background against which Congress legislate[s]."[268] As the *Babb* Court itself stated in the first sentence of its decision a few weeks earlier in *Comcast*, "[f]ew legal principles are better established than the rule requiring a plaintiff to establish causation."[269] In *Babb*, however, the Court effectively sanctioned a finding of liability without causation.[270]

Relatedly, as I argued above, and as Justice Thomas highlighted in his dissenting opinion, the any-consideration standard can be satisfied even if a plaintiff obtained a *favorable* employment outcome.[271] So long as age is considered in a promotion decision, an employee could satisfy the Court's standard, even if he in fact received the promotion.

Finally, like the motivating-factor test, the Court's "any-consideration" standard is vague. As discussed earlier in relation to the motivating-factor test, it is unclear what evidence would allow a jury to conclude that this standard is satisfied.[272] For example, would an earlier racist remark suffice? As in *Babb*, disparate-treatment cases frequently do not involve an explicit calculation, such as that provided in the Court's hypothetical. Ultimately, a jury would decide based on its intuition regarding *responsibility*, not causation. This vagueness can cause poor deterrence and inappropriate findings of liability, ultimately leading to suboptimal incentives and behavior.

The factorial framework does not encounter these problems. It applies a standard of causation that is based on the common and scientific notion of cause and effect. Unlike the standard articulated in *Babb*, causation under the proposed framework is determined with reference to the employment decision and is dependent on a form of difference in the employment *outcome*. Moreover, it is consistent with good policy and Congress's likely intent. In particular, it allows a finding of liability, even when discriminatory conduct is accompanied by other sufficient causes. Applying the NESS test to the Supreme Court's illustration, the test might ask, assuming the 35-year-old and 55-year-old had the same amount of points, would age have made the difference between the plaintiff receiving the promotion and not receiving

[267] Arguably, discrimination, or "differential treatment," can be understood as the "harm" in the Supreme Court's any-consideration framework. *See Babb*, 140 S. Ct. at 1178. But this reasoning similarly negates the meaning of causation by defining the harm in terms of the bad act itself rather than an outcome variable related to a potential consequence of the bad act.

[268] *Id.* at 1179 (Thomas, J., dissenting) (alteration in original) (quoting Univ. of Tex. Sw. Med. Ctr. v. Nassar, 570 U.S. 338, 347 (2013)).

[269] Comcast Corp. v. Nat'l Ass'n of Afr. Am.-Owned Media, 140 S. Ct. 1009, 1013 (2020).

[270] *See generally supra* Section III.A; *Counterfactual Causation*, *supra* note 15, at 887–93.

[271] *See Babb*, 140 S. Ct. at 1179 (Thomas, J., dissenting); *supra* Section III.A.

[272] *See, e.g.*, *supra* Section III.A.

the promotion? If he would receive a 5-point deduction due to age, then the NESS test is met.

On the other hand, the proposed framework excludes counterintuitive and counterproductive outcomes that could result from the vague any-consideration or motivating-factor tests. For example, it would exclude liability when discrimination is altogether preempted by another condition, such as the absence of minimal qualifications (e.g., the absence of a driver's license for a driving position) or the absence of a job opening. It would also not easily allow a racist jury to render a finding of no liability even when discrimination is a clear cause of an adverse employment decision.

Moreover, the *Comcast* and *Babb* decisions highlight two other advantages of the factorial framework. First, it is important to realize that *Babb*'s any-consideration test, like *Price Waterhouse*'s burden-shifting paradigm, severely limits a plaintiff's possible relief when the plaintiff cannot demonstrate but-for causation. In many contexts, this may lead to suboptimal incentives with respect to litigation and deterrence. For example, in *Babb*'s point-system illustration, it is possible that a plaintiff's inability to obtain monetary relief may undermine the incentive to sue, leading to suboptimal levels of deterrence.[273] Arguably, this is again a byproduct of choosing between two inadequate standards.

The factorial framework allows for a more nuanced approach to relief. If the protected feature satisfies NESS, then monetary relief could be available to the plaintiff, subject to a damages analysis, which may result in an apportionment of damages based on causation and other considerations.[274] This approach may promote better incentives.

Second, these two cases, decided only weeks apart, resulted in two distinct standards of causation. Arguably, the Supreme Court's decision in *Babb* involves an altogether *new* causal standard—the any-consideration test. The diversity of causal standards that govern disparate-treatment cases is arguably illogical and inequitable, but it is a natural consequence of choosing between two inadequate standards based on distinct language across various antidiscrimination statutes. The factorial framework avoids this problem. By providing a theory and standard of causation that is consistent with the causal language of the various statutes, as well as good policy and Congress's likely intent, the factorial approach allows a single causal framework that reflects actual cause and effect, and that is applicable to all disparate-treatment cases.

## VI. Conclusion

Antidiscrimination law's current approach to causation is deeply flawed. Its defects have caused an illogical, vague, and unworkable proof scheme that

[273] This is an open question that I do not attempt to resolve here.

[274] *See supra* Section IV.C.

requires an overhaul in order to curb the harm that it engenders and to allow the antidiscrimination statutes to serve their objectives effectively.

This Article proposes a causal theory and method that achieves this goal. In particular, the proposed factorial approach adopts a concrete and well-established scientific framework, the potential-outcomes framework, that permits clear and structured reasoning regarding causation, and an estimand, the NESS estimand, that refines and, in a sense, unifies the but-for and motivating-factor tests. By instilling nuance in the causal inquiry, and a logical causal estimand, the factorial framework permits a simple and effective method that is grounded in notions of actual cause and effect and basic tort law.

The factorial framework carries important implications for the policy objectives underlying antidiscrimination law. It employs a theory of causation that reflects actual cause and effect and that promotes the law's deterrence objectives while preventing windfall recoveries and their distorting effects on incentives. The factorial framework also allows an interpretation of causal language in antidiscrimination statutes that is consistent with good policy and Congress's likely intent. This is not possible if courts only consider the but-for and motivating-factor tests.

Finally, it is important to emphasize that the fact patterns underlying disparate-treatment cases are diverse, and this Article does not attempt to analyze the implications of the factorial framework with respect to the particulars of each and every possible disparate-treatment scenario. Additional nuance is surely needed. This Article, however, aims to establish a foundation for future analysis and development, and, ultimately, for the proposed framework's application as a logical and effective standard of causation in antidiscrimination law, and disparate-treatment cases in particular. This foundation could eliminate confusion and promote a more coherent and effective system of antidiscrimination law.