Introduction

Social scientists ask diverse kinds of research questions. Usually, each such question calls for application of a specific analytic strategy to empirical evidence. For example, questions about the distribution of wealth in a population call for the analysis of variation in levels of wealth across a sample of households, using sociodemographic and other variables to predict levels. Analytic methods for the study of distributions are especially well developed in the social sciences today. Variation in a dependent variable (e.g., household wealth) is explained using variation in independent variables (e.g., race, ethnicity, immigration status, education). Social scientists have developed a vast array of variation-based analytic techniques, perfect for addressing questions about distributions.

But not all research questions are so lucky. Often, the research goal is to understand "how" a qualitative outcome happens by examining a set of cases that display the outcome. The distribution of that outcome in a sample drawn from a population will be relevant, but the empirical focus in determining the *how* of the outcome must rest on cases that display the outcome. Cases without the outcome—key evidence in the analysis of variation in the distribution of the outcome—can provide only very limited information regarding how the outcome happens. Restricting the analytic focus to cases that display the outcome, however, transforms the "dependent variable" into a *constant*—which precludes using the many variation-based analytic techniques that social scientists have developed. There is no readymade technique, comparable in sophistication to techniques that rely on a dependent *variable*, for the analysis of constants as outcomes.

Questions regarding how outcomes happen are quite common, though—especially in everyday discourse. Unfortunately, they are often recast by social scientists as questions about distributions. Imagine, for example, that instead of learning about the process of becoming a marijuana user by observing and interviewing users, Howard Becker (1953, 2015) had instead examined the distribution of marijuana use in a random sample drawn from a given population. Suppose he

found high levels of use among musicians and certain other, related groups. While indirectly relevant to the *how* question, the finding does not address it head-on. To find out how one becomes a marijuana user, it is necessary to study users, focusing especially on their shared experiences in learning to use marijuana and on other widely shared antecedent conditions.

This book offers a straightforward methodology for the assessment of research questions regarding the antecedent conditions linked to qualitative outcomes. A typical qualitative study has a set of cases that display the outcome in question—the *focal* outcome—along with evidence on relevant antecedent conditions. The goal of the analysis is to identify antecedent conditions shared by cases with the focal outcome. Shared antecedent conditions, in turn, may be interpreted as "recipes" for an outcome, especially when they make sense as combinations of causally relevant conditions. In the end, the researcher explains a constant (the focal outcome) by way of other constants or near-constants (shared antecedent conditions).[1]

My approach to the analysis of systematic cross-case evidence on qualitative outcomes has deep roots in sociology in the form of a technique known as analytic induction (AI). AI was a popular research technique in the early decades of empirical sociology, beginning with the publication of Florian Znaniecki's (1934) *The Method of Sociology* (Tacq 2007). Exemplary AI studies include Alfred Lindesmith's (1947, 1968) *Addiction and Opiates*, Donald Cressey's (1953, 1973) *Other People's Money*, and Howard Becker's (1953, 2015) *Becoming a Marihuana User*. AI seeks to establish invariant (or "universal") conditions for qualitative outcomes, focusing exclusively on instances of the outcome and how it came about in each case.

As explained in chapter 1, early applications of AI used an especially strict version of the approach, which I call "classic AI." Classic AI (see also Becker 1998: 196–97) is strict in that it does not permit disconfirming cases, defined as cases where the outcome is present but one or more of the antecedent conditions specified in a working hypothesis is absent.[2] All instances of the outcome must be accounted for in some way, either by narrowing the definition of the outcome, thereby excluding disconfirming cases, or by respecifying the relevant antecedent conditions in a way that accommodates the disconfirming cases (see chapter 2). In fact, disconfirming cases are essential to classic AI because they provide raw material for refining the researcher's working hypothesis. They push the analysis forward. Very often, classic AI researchers seek out disconfirming cases, in order to refine their arguments, and in this way AI is akin to grounded theory's utilization of theoretical sampling based on inductively derived categories (Glaser and Strauss 1967; Katz 2001; Hammersley 2010).

However, as is so often the case with analytic methods, classic AI's strength is also its weakness. Accounting for every disconfirming case, as defined above, requires both in-depth knowledge of cases and substantial conceptual agility on the part of the researcher (see chapters 2 and 3). Besides, social phenomena are

both heterogeneous and chaotic, data collection methods are imperfect, measures are crude and often contain known or hidden biases, revisits to research sites or subjects are often difficult or impossible, and coding mistakes are all too common (Katz 1983). One researcher's coding error is another researcher's disconfirming case, just as one ethnographer's observation of a wink is another ethnographer's observation of a blink. In principle, addressing disconfirming cases is a great way to fine tune a working hypothesis; in practice, however, it is often difficult to achieve satisfactory results (Becker 1958; Bloor 1978; Katz 1983).

Consequently, systematic applications of classic AI today are relatively rare. Instead, researchers interested in systematic cross-case evidence on qualitative outcomes routinely construct what I like to call *composite portraits* of their cases. For example, a researcher interested in the process of becoming a committed social movement activist might collect interview data on a diverse set of committed activists and attempt to identify common background characteristics and other shared antecedent conditions (see, e.g., Downton and Wehr 1998; Driscoll 2018). The researcher in this example would not expect to find every important background characteristic in every activist—as required by classic AI. Instead, the goal would be to identify background conditions that are widely shared by activists. The end product in this example would be an idealized composite portrait—an "ideal typic" (Weber 1949) activist who combines the major background characteristics identified by the researcher.

The composite portrait approach, as just described, has a lot in common with classic AI. The analytic scope is limited to cases that display the focal outcome. The research question asks, "How did the outcome happen, or come about?" The focus is on widely shared antecedent conditions, the expectation is that there are multiple antecedent conditions, and the researcher's goal is to make sense of shared conditions as a formula or recipe for the focal outcome. In fact, the pivotal difference between classic AI and the composite portrait approach just described is classic AI's insistence on identifying *invariant* antecedent conditions. For these reasons, it is appropriate to refer to the composite portrait approach as "generalized AI." It is generalized in the sense that it is a flexible adaptation of AI to the chaotic and capricious nature of social phenomena and to the many practical challenges of establishing invariant relationships.

As a substitute for classic AI's invariance requirement, generalized AI attends to frequency criteria. That is, the researcher attempts to identify *widely shared* antecedent conditions, not universally shared conditions. Thus, "enumerative" criteria—simple counts and proportions, for example—are utilized, but they are used to gauge the consistency of antecedent conditions, not to assess bivariate or multivariate relationships (Goertz and Haggard 2022). The latter would require an outcome that varies across cases, which AI eschews. Evaluating the generality of antecedent conditions across a range of positive cases—generalized AI's core procedure—is essentially an assessment of the "consistency" of set-theoretic relations

TABLE I-1 Contrasts between generalized analytic induction and conventional variable-oriented research

	Generalized analytic induction	Conventional variable-oriented research
Outcome	Constant across cases	Varies across cases
Focus	Causal formula or "recipe" based on shared antecedent conditions	Net effects of independent variables on a dependent variable
Scope of analysis	Cases with the outcome	A given population or defined set of candidates for the outcome
Negative cases	Not directly relevant	Essential
Explanatory template	Constants explain constants	Variables explain variables
Case selection	Diverse set of instances of the outcome	Representative sample drawn from a population or defined set
Research question	How the outcome happens	Relative effects of independent variables on the distribution of an outcome

(Ragin 2008: chaps. 1–3). Thus, generalized AI is best understood as a set-analytic technique, not a correlational one.

As an approach to social research, generalized AI differs fundamentally from conventional, variation-based approaches. The important contrasts between the two approaches are summarized in table I-1. As noted previously, generalized AI's outcome is a constant—the set of cases displaying the outcome in question. While most such outcomes are qualitative in nature, it is possible as well to base the analysis on cases that meet a specified threshold of a quantitative variable (e.g., an income level signaling that an individual is well-off—see chapter 9). Conventional variable-oriented research, by contrast, is centered on the task of explaining variation in a dependent variable, focusing on the net effects of independent variables (Ragin 2006b). Another key contrast is the role of "negative" cases—that is, cases that fail to exhibit the focal outcome. Such cases are not considered disconfirming according to generalized AI. Instead they are considered instances of an alternate outcome and therefore are the focus of a separate analysis altogether. By contrast, negative cases in conventional quantitative research are valued for their contribution to variation in the dependent variable.

It is important to point out that unlike much variable-oriented research, generalized AI is not inferential. Instead, it is primarily descriptive and is best understood as an aid to causal interpretation. It can be used in conjunction with other analytic methods, including conventional quantitative methods, by providing results in the form of causal recipes. Conventional quantitative methods focus primarily on isolating the separate, net effects of "independent" variables, not on their conjunctural impact. This aspect undermines the utility of conventional

quantitative methods for causal interpretation, which often involves a focus on recipe-like combinations of conditions.

The application of generalized AI's core procedure is ubiquitous in social research, especially in qualitative work (Bernard et al. 2017). It's obvious that a lot can be learned from exploring the antecedent conditions shared by positive instances of an outcome (Goertz and Haggard 2022). Unfortunately, most applications of the core procedure are unsystematic and ad hoc. Only rarely do researchers quantify their assessments, and seldom do they explore *combinations* of conditions linked to an outcome. My main argument in this book is that there is a lot to be gained from systematizing generalized AI as a set-analytic method. In the chapters that follow, I make the case for treating generalized AI as a formal technique (see also Ragin and Amoroso 2019: 112–17).

OVERVIEW

Part I of this book (chapters 1–4) examines classic AI and addresses basic research-design issues associated with its use. Chapter 1 introduces the method, detailing its logic, describing it as a series of steps, and reviewing some exemplary applications. I also touch on the controversy stirred by classic AI, especially following W. S. Robinson's (1951) critique in the *American Sociological Review*. Along the way, I compare correlational approaches to causation with set-analytic approaches and describe AI's contrasting approach to two very different kinds of "disconfirming" cases: those that display the antecedent conditions specified in a working hypothesis but not the outcome, and those that display the outcome but not the hypothesized antecedent conditions.

Chapter 2 offers a thorough discussion of AI-based methods for addressing disconfirming cases—that is, instances of an outcome that fail to display the antecedent conditions specified in the researcher's working hypothesis. There are two main strategies for reconciling such cases. One is to narrow the definition of the outcome so that disconfirming cases are excluded. The other is to expand the breadth of the working hypothesis in a way that accommodates the disconfirming cases. It is also possible to address disconfirming cases by developing outcome subtypes or through the specification of appropriate scope conditions.

Chapter 3 examines the methodological implications of two very different types of research questions. On the one hand, what explains variation in the level or probability of an outcome? On the other, what explains the focal outcome's occurrence—how it comes about? The key is that the first question is focused on the distribution of an outcome in a given sample or population, while the second is focused more or less exclusively on positive instances of the outcome. These two different ways of conducting social science have spawned widespread disagreement and controversy. In one camp, researchers who seek to explain variation reject the other side's "selection on the dependent variable." Meanwhile, in the opposing

camp, researchers focused on understanding how instances of an outcome happen reject a common practice of the other side: boosting the sample size of cases by casting a wide net, thereby running the risk of including irrelevant cases.

Chapter 4 contrasts three approaches to the analysis of dichotomous outcomes: conventional quantitative analysis, qualitative comparative analysis (QCA), and AI.[3] The three approaches can be arrayed along a continuum with respect to the dependence of standard applications of each approach on the analytic incorporation of "negative" cases. Conventional quantitative analysis is fully dependent on negative cases, and its treatment of negative cases is fully symmetrical with its treatment of positive cases. Most applications of the second approach, QCA, are also dependent on negative cases, but in a different manner. QCA's truth table procedure uses negative cases to classify truth table rows as true or false based on the degree to which the cases in each row consistently display a given outcome. By contrast, negative cases of the outcome play no direct role in AI, which separates the analysis of positive cases from the analysis of negative cases. In this "fully asymmetric" approach, negative cases are viewed as positive cases of one or more alternate outcomes.

Part II (chapters 5–10) offers a detailed presentation of generalized AI. Chapter 5 introduces Part II by briefly summarizing key differences between generalized AI and classic AI. Chapter 6 describes an essential feature of generalized AI: its reliance on "interpretive inferences" based on substantive and theoretical knowledge. Interpretive inferences transform presence-versus-absence conditions into contributing-versus-irrelevant conditions. For example, substantive knowledge indicates that being educated contributes to avoidance of poverty. On the basis of this knowledge, a researcher would bypass consideration of "not being educated" as a condition for avoiding poverty. If a person who has successfully avoided poverty is uneducated, then their lack of education is eliminated as a possible contributing condition of their avoidance of poverty. This feature of AI contrasts sharply with QCA's configurational logic, which requires both sides of every presence/absence condition to be treated equally.[4] Configurational logic dictates that the researcher entertain the possibility that not being educated could contribute to successfully avoiding poverty.

Using hypothetical data on Olympic-caliber athletes, chapter 7 offers a step-by-step application of generalized AI to the analysis of a set of cases that share the outcome "sustained commitment." Many researchers, especially those who conduct qualitative investigations, are routinely tasked with making sense of a set of instances of an outcome. Because the outcome in question does not vary, a conventional quantitative approach is of little use here—as is, without negative cases, QCA (as demonstrated in chapter 6). By contrast, generalized AI provides important tools for making sense of such cases.

A reanalysis of data published in Jocelyn Viterna's (2006) study of women's mobilization into the Salvadoran guerrilla army is the focus of chapter 8. Viterna

applies key principles of generalized AI in her pathbreaking study. She distinguishes five different outcomes—three distinct paths to guerrilla activism (politicized, reluctant, and recruited) and two non-guerrilla paths (collaborators and nonparticipants). Rather than define the analysis as a binary contrast between the three guerrilla paths versus the two non-guerrilla paths, she focuses instead on the separate conditions linked to each of the five outcomes. She views each of the outcomes as worthy of separate analytic attention and thereby avoids conventional dichotomization of the outcome as "guerrilla versus non-guerrilla." This feature of her study, along with several others, aligns well with generalized AI.

Chapter 9 tackles the problem of bridging generalized AI and conventional quantitative analysis. It demonstrates that generalized AI can be usefully applied to conventional quantitative data. Because generalized AI is fundamentally descriptive in nature, it can complement findings derived using conventional quantitative methods. The demonstration of generalized AI uses data on Black females from the National Longitudinal Survey of Youth (NLSY), 1979 sample. The focus is on two outcomes, analyzed separately: membership in the set of respondents in poverty, and membership in the set of respondents well out of poverty. The results are asymmetric, with different conditions linked to the two outcomes.

The final chapter summarizes the essential features of generalized AI, as presented in this book. The listed features range from generalized AI's orientation as a research approach to practical procedures involved in applying the method.

A NOTE ON THE CONCEPT OF CAUSATION

The primary objective of this book is to provide tools that aid causal *interpretation*. Tools for causal *inference*, by contrast, are beyond its scope. More generally, the approach to causation advocated in this work is based on the regularity theory of causation. According to this theory, causation is indicated by an invariant connection between cause and outcome, which is also a concern of classic AI as described in this book. Classic AI adheres to John Stuart Mill's version of regularity theory, in particular his method of agreement, which selects on instances of an outcome and seeks to identify their shared antecedent conditions.

The relation between antecedent conditions and outcomes is set-theoretic in nature: instances of the outcome constitute a subset of instances of the antecedent conditions. This subset relation is evident, for example, whenever instances of an outcome agree in sharing a causally relevant antecedent condition. Of course, perfect set relations are relatively rare in social research. Thus, this book emphasizes assessing the degree of consistency of empirical evidence with the subset relation in question and restricts the analytic focus to connections that are highly consistent.

The designation of conditions as *causally relevant* to an outcome is dependent on theory and knowledge, and thus open to contestation. The larger task of specifying the "true" causes of social phenomena is beyond the scope of this work, and

indeed beyond the purview of most social science methodology. Usually, social scientists must be content with successfully identifying causally relevant antecedent conditions, which in turn are suggestive of causal mechanisms. The true test of any hypothesized antecedent condition is its relevance at the case level. It is at the case level that social researchers have the opportunity to observe and narrate causal processes and mechanisms (Goertz and Haggard 2022). Thus, establishing regularities is essential, but it is not the whole story. Whenever possible, researchers should complement the identification of regularities with confirmatory process tracing at the case level.