1

The Diversity of Views on Causality and Mechanisms

This chapter is devoted to three conceptual clarifications. First, I clarify the meaning I give the concept of causal inference throughout the book (Section 1.1); second, I discuss the concept of causality (Section 1.2); and, finally, I investigate the concept of mechanism (Section 1.3). An explicit discussion of the variety of meanings the concepts of causality and mechanism usually receive in different academic communities is useful to highlight an interesting fact: specific views on causality tend to square with specific views on mechanisms. Once this is clearly seen, it will become apparent that the observed disagreements on the usefulness of an agent-based model (ABM) for causal inference ultimately arise from the fact that scholars in different methodological traditions endorse conflicting views on what establishing causality and identifying mechanisms mean (Section 1.4). To establish this fact is the first step, I believe, to building a pluralistic and comprehensive view on causal inference in quantitative social sciences.[1]

1.1 Causal Inference

Historically, statistical inference was created to solve the fundamental problem of gaining knowledge about given features, i.e. parameters, of a given population by observing only a portion of it, i.e. the sample (see Goldthorpe 2021: chs. 3–4)? Probability theory—in its frequentist and subjective variants (see Cox 2006)—was proposed as the tool for quantification of the errors that one inevitably makes when studying the parts rather than the whole.

Once this mathematical invention entered social sciences, statistical inference was stretched in two different, but related, ways. On the one hand, as described by Ziliak and McCloskey (2008), social scientists have progressively tended to give priority to establish whether parameters, and their association, are different from zero rather than focusing on estimation of their value. Thus, the question of the existence of an effect, rather than its size, becomes the main goal of statistical inference, and statistical tests, with their significance values, the main tool to make a decision. On the other hand, as noted by Freedman (1995), many social scientists were incapable of resisting the temptation of considering a presumably different-from-zero association between two features as the sign of the existence of a

[1] This chapter builds on and extends Casini and Manzo (2016: 6–14).

Agent-based Models and Causal Inference, First Edition. Gianluca Manzo.

causal connection between them. Thus, various forms of regression analysis became the main tools to establish causal claim, and statistical inference was silently equated to causal inference. In this way, Freedman (2005) notes, a second interpretative layer was added. The same set of assumptions that was initially created to determine whether, under a specific theory of (a specific type of) errors, a feature had a certain value in the population was now extended to the interpretative, not factual, task to make this value contingent on the estimated value of another feature. The shift that was progressively operated in quantitative social science from statistical inference to causal inference is nowadays especially visible within the literature on the potential outcome approach as a comprehensive framework to interpret both experiments and multivariate statistics for observational data (see, in sociology, Morgan and Winship 2015; in economics, Imbens and Rubin 2015).

Throughout this book, I adopt a generic definition of causal inference, which I regard as the cognitive operation that consists in using fragmentary empirical and theoretical information to establish the extent to which one specific happening systematically alters the probability that another happening follows. This definition is motivated by the theory of "epistemic causality" (see Williamson 2006) and the "evidential variety" thesis (see Russo and Williamson 2007), which consider that defending a causal claim always requires an investigator and an audience: the investigator's goal is to accumulate empirical data and theoretical arguments that must be as diverse as possible to make her causal claim as persuasive as possible in the eyes of the given audience. From this perspective, it is unlikely that the investigator is able to persuade the audience that her causal claim is plausible as long as she only produces a single type of evidence originating from a single type of method. As a consequence, there is no reason to restrict *ab initio* the concept of causal inference to a specific type of method and evidence, namely statistical methods for survey data or experimental designs. This open view of what counts as evidence for causal inference echoes some statisticians' call for recognizing that confidence in statistical methods and statistical results is always built in practice at the intersection of a complex mix of elements of different natures, which suggests problematizing what "convincing evidence" is in practice (see, in particular, Gelman and O'Rourke 2013; Gelman and Basbøll 2014).

Consistently with this view, let me emphasize that, although I will argue that generating the connection of interest from formal models of mechanisms through ABMs constitutes the "severe test" for making persuasive causal claims (on "severe testing", see Mayo 2018), I restrain myself from defining causal inference in general in terms of mechanism-based reasoning (for a different choice, see Hedström 2009). In my view, distinguishing analytically the cognitive operation of inferring causality from the specific types of arguments and methods one exploits to perform this operation is the very preliminary conceptual step we need to appreciate precisely that a variety of methods, data sources, and theoretical arguments actually should be combined to make persuasive causal claims.[2]

---

[2] This seems the right point of the analysis to make an important linguistic observation. From the very beginning of the book, I tried to be explicit about one of the major points I want to make, namely the idea that assumptions are a crucial part of any causal reasoning, which implies that "evidence" can never be conclusive *alone* because it is always conditional on those assumptions. Despite this starting point, in a previous version of the manuscript I recurrently adopted the expressions "convincing evidence" and "convincing causal claims". Christopher Winship made me realize that this linguistic choice may have

1.2 *Dependence* and *Production* Accounts of Causality

To argue in favor of a pluralistic perspective on causal inference, the first step is fully to acknowledge that it is hard to restrict the concept of cause to a single, and uncontroversial, meaning. The point is nicely illustrated by Nancy Cartwright (2004: 806, italics mine), when she notes that "[t]he term cause is highly unspecific. It commits us to nothing about the kind of causality involved nor about how the causes operate. Recognizing this should make us *more cautious* about investing in the quest for *universal methods for causal inference*."

This is an important epistemological observation that has the following implication. If the concept of causality can theoretically be understood in different ways, and if different quantitative methods are differently permeable to different theories of causality, then it is likely that one's view on what method is best suited to perform causal inference in fact depends on one's favored intuitions on what causality is or is not. Thus, making explicit such intuitions is especially important for assessing in what sense and under which conditions methods that are usually excluded from the causal business—like ABM—can contribute to causal inference *compared* to more well-established experimental and observational methods.

Contrary to the skepticism of many quantitative social scientists against philosophers of causality[3] I do believe that philosophical writings on causation provide theoretical coordinates to map the different ways we can conceive of a causal relationship. In particular, for my purpose, the most relevant point is the generic distinction between *dependence* (or *difference-making*) accounts of causality and *production* accounts of causality (see Hall 2004). Roughly, among dependence accounts, one finds regularity, probabilistic and counterfactual views of causality whereas, among production accounts, one finds process, entities-andactivities and dispositionalist theories (for similar categorizations, see Kistler 2002; Psillos 2007; Reiss 2013: ch. 5). Obviously subtle differences between each variant of the two perspectives exist. For my argument, however, what matters most are the basic intuitions that inspire these two groups of theories of causation.

In particular, the idea behind dependence accounts is that causes are such that their obtaining makes a difference to the obtaining of their effects. In contrast, the idea behind

---

run against my own central argument because the adjective "convincing" may convey the impression that, despite all my qualifications, there can be in some sense some forms of absolutely convincing evidence (or at least that I believe that this is ultimately possible). To avoid this misunderstanding, I substituted "convincing" for "persuasive" (I must thank Winship again for suggesting this alternative), and limited the use of the first term to statements where no ambiguity seemed possible to me. For the same reason, I will very often write "data and arguments" rather than "evidence" with the intent of being entirely explicit about the point that the kind of support that any method can provide for a given causal claim is always a combination of partial empirical information and arguments that are necessary to defend assumptions that are not, or are only partially, empirically verifiable.

[3] For instance, Freese and Kevern (2013: 28) claimed: "The professional philosophical literature on causality is often surprisingly unhelpful: the practically-minded researcher digs in looking for clarity and instead is soon invited to consider examples of simultaneous assassination attempts or billiards rolled into time machines. No uncontroversial general philosophical account of causality exists, and social researchers have plenty of our own work to do while we wait".

production accounts is that causes are such that they help generate, or bring about, their effects. To Hall (2004: 226), from whom I am taking this conceptual distinction, dependence and production accounts are irreducible to one another, so we have distinct concepts of cause. In the present investigation, I remain agnostic on what causality is ontologically. My ultimate goal is to propose a comprehensive and pluralistic framework for causal inference that can accommodate several types of methods, sources of data, and theoretical arguments. For this reason, I am reluctant to accept that distinct and irreducible notions of cause exist. On a methodological level, this view has undesirable consequences, which I will discuss later (see Chapter 6).

For now, let me emphasize that, independently from philosophers of science, and under different terminological labels, sociologists have developed categorizations that follow Hall's distinction between dependence and production accounts of causality. In particular, more explicitly than others, Goldthorpe (2001) remarked that causation can be interpreted at least in three different ways as "robust dependence", as "consequential manipulation", or as "generative process". According to him, in the first case, "the causal claim depends on showing that X continues to affect Y when a set Z of other variables, also possibly related to Y, are introduced in the analysis" (*ibid*.: 2). In the second case, "genuine causation is that if a causal factor, X, is manipulated, then, given appropriate controls, a systematic effect is produced on the response variable, Y" (*ibid*.: 4–5). Finally, when causation is understood as a "generative process", "(...) what is important is the nature and the validity of the account given of the process that underlies the association appealed to (...)" (*ibid*.: 9). In terms of the aforementioned philosophical categories, causation as "robust dependence" and "consequential manipulation" clearly exemplify *dependence* accounts of causality, whereas what Goldthorpe labels causality as "generative process" naturally falls within *production* accounts.

Goldthorpe also observed that these views on causality combine in practice with distinctive methods of social inquiry. According to him, the view that causality essentially depends on controlling for confounders has informed time-series analysis, the early generation of causal path analysis, structural equations models, and more generally the large panoply of multivariate quantitative, regression-like techniques for survey data analysis (on this point, see also Abbott 1998; Freedman 2005). The "consequential manipulation" view squares with the methodology of randomized experiments. Interestingly, Goldthorpe hesitates to identify a specific method that illustrates the view of causality as a "generative process". Although he sees the potential of simulation methods as a possible option to test the validity of a proposed account of the underlying process (*ibid*.: 14), he still prioritizes statistical methods with respect to the goal of testing the hypothesized direct and indirect consequences of a postulated underlying process (*ibid*.: 12–3)—a view that he restated in his manifesto for sociology as population science (see Goldthorpe 2016: ch. 9), where, however, differently from early writings, he explicitly mentioned "agent-based computational modeling" as a possible methodological option for studying models of social mechanisms.

Interestingly, Ermakoff (2019) found, inductively, similar conceptual distinctions and methodological associations by reviewing scholarship in comparative and historical sociology. In particular, he observed that three different types of causal investigation are at work in this field: (i) a "morphological" approach, which bases causal claims on the detection of spatial and temporal patterns of social phenomena by relying on various

descriptive techniques of data reduction (*ibid*.: 3–5); (ii) a "variable-centered" mode of causal analysis that infers causal claims "from patterns of association among a set of empirical categories" (*ibid*.: 6) by relying on multivariate statistical analyses that "probe the statistical significance of correlations and imputed effects" and comparative analyses detecting "combinations of attributes across cases"; and (iii) a "genetic" approach that "apprehends causality through the systematic investigation of generative processes" (*ibid*.: 11), a mode of investigation that Ermakoff connects to various methodological approaches to theorize and validate models of mechanisms, among which are "agent-based simulations" (*ibid*.: 16). Thus, on the one hand, what Ermakoff calls "morphological" and "variable-centered approach" squares with Goldthorpe's accounts of causality as "robust dependence" and "consequential manipulation"—Ermakoff explicitly evokes "the counterfactual framework" as a tool for "causal diagnoses" within the "variable-centered" approach (*ibid*.: 10)—and, on the other hand, Ermakoff's "genetic" category precisely correspond to Goldthorpe's type of causation as "generative process".

The association between specific methods for causal inference and views on causation—an association recently also documented by Kaidesoja (2021a)—is especially visible in the recent and flourishing literature on the so-called "potential outcome" approach. This is driven by the ambition to introduce the perspective of randomized experiments into the analysis of data generated outside an experimental setting (for a historical overview, see Imbens and Rubin 2015: ch. 2). Accordingly, the main task of the analysis becomes to show that individuals (or other units of analysis) that are exposed to different treatment states are likely to exhibit different responses, or outcomes. The causal effect of a given treatment state is then conceived as the (average) difference between the outcome of those who were exposed to it and that of those who were not.

In this way, the potential outcome approach is in essence tied to a counterfactual understanding of causation, which, as I have noted above, falls within the *dependence* (or *difference-making*) account of causation. From this perspective, establishing causal claims indeed amounts to quantify what-if outcomes, i.e. how a given group of units of analysis would have responded had their treatment value been different. As noted by Morgan and Winship (2015: 4), what-if (or potential) outcomes are counterfactual in the sense that they "exist in theory but are not observed". This is an important observation. It implies that, by construction, establishing causal claims within the counterfactual approach cannot be seen as a pure matter of empirical data. I will develop this line of reasoning further in Chapter 5 (see Section 5.3).

For the moment, the important point here is that the potential outcome approach, with its counterfactual understanding of causation, is now regarded by many as a "unified framework for the prosecution of causal questions" (Morgan and Winship 2015: 3). As such, it is seen as a tool that allows one to recast traditional multivariate statistical instruments in the terms of this particular view of causality. In this regard, Morgan and Winship's discussion of matching and regression estimators (2014: chs. 5–7) is especially illuminating. They elegantly show how the classic method of controlling for confounders can be reinterpreted as aiming not so much to identify "robust dependences"—to go back to Goldthorpe's distinctions—as to render comparable the outcomes of group subjects that were not randomly assigned to the treatment state of interest (see also Hernán and Robins 2020: ch. 15).

The diffusion of the potential outcome approach had an additional important consequence. It made explicit a conceptual distinction that helps to recognize that scholars understanding causality in terms of dependence (rather than production) relationships can themselves raise different causal inference questions. The distinction is that between what Gelman (2011: 955) proposed to call "forward" and "reverse" causal inference. According to him, when one is pursuing "forward" causal questions, one seeks to understand and quantify "what may happen if we do X" whereas, when one is interested in "reverse" causation, one wants to answer the question of "what causes Y". In the former case, one focuses on one specific phenomenon (education, for instance) and wants to establish the consequences of its presence, absence or variation (on fertility, for instance) whereas, in the latter case, one observes a given outcome and *a posteriori* wants to trace back the outcome to the various phenomena that may have made it happen. For this reason, by using a terminology already present in John Stuart Mill's (1882) *A System of Logic, Ratiocinative and Inductive* (see in particular chs. 6, 7, and 10), the expressions "effects of causes" and "causes of effects" are also now often used to refer to "forward" and "reverse" causation respectively—the latter in fact being often also called "backward" causation (see, for instance, Sampson et al. 2013: 3, 7, 24). The point I want to stress here is that, as Gelman (2011: 956) acknowledged, the potential outcome approach is precisely the framework used by statisticians, and many economists, to treat "forward" or effect-of-a-cause questions, and experiments are seen as the prototypical method to address this type of causal inference question (see also Dawid et al. 2014). In contrast, many social scientists continue to address "reverse" or "cause-of-effect" questions through multivariate statistical techniques, a choice that the potential outcome perspective invites to question in terms of the assumptions that are needed to establish the "causal" nature of each "cause" and its relative weight (see Gelman and Imbens 2013). Interestingly, this contrast echoes an old intuition that Mill (1882: 557) already expressed in the following way: "since, as a general rule, the effects of causes are far more accessible to our study than the causes of effects, it is natural to think that this method has a much better chance of proving successful than the former". The method Mill was referring to precisely amounted to intervene on, or manipulate, experimentally the factor of interest.[4]

Now, although it seems unquestionable that the potential outcome approach to causal inference, and its associated effect-of-a-cause view, forces quantitative scholars to think more rigorously, and more modestly, about how statistical methods for observational data can be used to defend causal claims, it should be pointed out that the causality account embedded within the potential account approach is still highly specific. In terms of the aforementioned philosophical distinctions, the counterfactual view of causation associated with the potential outcome approach is a type of *dependence*, or *difference-making*, account of causality. From within a "production" perspective, this counterfactual view may be regarded as limited, in the sense that, in the words of the statistician

[4] I should thank Christopher Winship for pushing me to discuss explicitly the distinction between "forward" and "reverse" causal inference, and articulate it more clearly with my double distinction between dependence and production accounts of causality, on the one hand, and horizontal and vertical mechanisms, on the other hand (see, on this point, my remarks in Section 1.4).

David Cox (1992: 297), it lacks "an explicit notion of an underlying process or understanding at an observational level that is deeper than that involved in the data under immediate analysis". To this, Cox adds: "my preference, however, is to restrict the term [causality] to situations where some explanation in terms of a not totally hypothetical underlying process or mechanism is available". Remarkably, Cox's critique is reminiscent of oldest statements by the realist philosopher Rom Harré (1972: 115–9, 136–7), who, already in the 1970s, remarked that the "successionist" view of causality should be complemented with a "generative" theory of causality. In Harré's (1972: 137) words, "science is based upon the generative theory, and treats the statistical evidence of succession as the basis for the hypothesis that a causal mechanism exists. This generates a methodological principle, in that a study is deemed complete only when the causal mechanism has been identified (...)".

Thus, similarly to philosophers of science, sociologists and statisticians clearly subscribe to different accounts of causality (for a clear synthesis, see Brady 2011: table 49.1). These different accounts tend to come with equally clear qualitative judgment on the respective merits of the various accounts. Goldthorpe (2001: 8–9), for instance, clearly states that the view of causation as "generative process" should be seen as an improvement on the "robust dependence" and "consequential manipulation" accounts because "it would appear to derive, rather, from an attempt to spell out what must be added to any statistical criteria before an argument for causation can convincingly be made". Hedström (2009), similarly, remarks that only the presence of a fully fledged mechanism authorizes causal inference and allows one to reach explanatory depth. To be sure, scholars within the potential outcome tradition would find this priority judgment unjustified because, so they would claim, mechanism-based explanations can be easily formulated within a counterfactual view and tested by an appropriate use of statistical methods (see Morgan and Winship 2015: ch. 10). As I will show next, however, different understandings of the concept of mechanism are at work here, thus increasing further the probability of misunderstanding and miscommunication between scholars that already understand causality in different manners.

1.3 *Horizontal* and *Vertical* Accounts of Mechanisms

As we have seen, "production" accounts of causality—differently from "dependence" ones—require the identification of an underlying mechanism for inferring causality from data. But what is a mechanism exactly? Similarly to the concept of causality, the concept of mechanism, too, has received a variety of interpretations (in philosophy, see Reiss 2013: 104–5; in sociology, see Mahoney 2001: 579–80; Hedström 2005: 25; Gross 2009: 360–2; in political science, see Gerring 2008). As recently observed by Kalter and Kroneberg (2014), the term mechanism has clearly penetrated much empirical research in sociology but it is still employed with a variety of meanings. Mapping this variety is important because, depending on how a mechanism is understood, ABM will be seen as either necessary to study models of mechanisms or unnecessary, and, consequently, different judgments will be formulated on the potential contribution of ABM to causal inference.

Again philosophical scholarship on mechanisms provides useful theoretical coordinates to appreciate various views on mechanisms (for an overview, see Andersen 2014a, b). For my purposes, the most relevant distinction here is between what I propose to call the "*horizontal*" and "*vertical*" views of mechanisms.

According to the former view, a mechanism is interpreted as a network of variables that stand in particularly robust relations. Woodward (2002: S375) exemplifies this view when he defines a model of a mechanism as a description of "(...) (i) an organized or structured set of parts or components, where (ii) the behavior of each component is described by a generalization that is invariant under interventions, and where (iii) the generalizations governing each component are also independently changeable" (for a discussion of the "invariance" and "modularity" conditions, see Kaidesoja 2021b). In contrast, according to the *vertical* view, a mechanism is envisaged as a "complex system" (Glennan 2002: S344) comprising a set of unities—entities and activities (Machamer et al. 2000), or component parts and operations (Bechtel and Abrahamsen 2005), or parts and interactions (Glennan 2002)—that, by interacting over time, generate some behavior of the system. Machamer et al. (2000: 3) exemplify this view when they define a mechanism as "(...) composed of both entities (with their properties) and activities. The organization of these entities and activities determines the ways in which they produce the phenomenon."

Although the vertical view is not incompatible with paying attention to the robustness of the relationships between the interacting parts that compose the mechanism (for an elegant discussion of this complex nuance, see Baumgartner et al. 2020), the distinctive feature of a mechanism from a vertical perspective is the dynamic of the changes a mechanism brings about. From their activity-centered perspective, Machamer et al. (2000: 3) put this point by saying that "entities often must be appropriately located, structured, and oriented, and the activities in which they engage must have a temporal order, rate, and duration" and that "description of a mechanism describes the relevant entities, properties, and activities that link them together, showing how the actions at one stage affect and effect those at successive stages" (*ibid*.: 12). From his interaction-centered perspective, Glennan (2002: S344) makes the same point when he remarks that "[a] mechanism operates by the interaction of parts. An interaction is an occasion on which a change in a property of one part brings about a change in a property of another part." In a word, what matters to the vertical view is the sequence of micro-level changes that dynamically create new connections within the system under scrutiny.

Thus, through the choice of the terms *vertical* and *horizontal*, I primarily intend to grasp the idea that, within the vertical view of a mechanism, the supposedly causal connection of interest is seen as gradually created by combining progressively the small changes triggered by the activities and the interactions of the low-level entities underlying the connection of interest. From this perspective, a model of a mechanism is conceived as a detailed account of *how* one moves from the low (small) level (scale), whatever it is, to the higher (or larger) level (scale). Andersen (2014a) expressed a very similar idea when she noted that treatments of mechanisms in the style of Glennan or Machmaer are intrinsically "hierarchical": the emphasis is on how nested entities and activities create connections between different levels of analysis. In contrast, within the horizontal view of a mechanism, the goal of providing a granular account of the changes that bridge levels of analysis

(or different scales) is put into parenthesis (on granularity, see Chapter 2 and Chapter 4, Section 4.2.2). Andersen (2014b: 286) formulated this idea by noting that accounts of mechanisms *à la* Woodward are intrinsically "flat" in the sense that "the variables are all at similar levels (of size, organization, or other level differentiation)".[5]

With these qualifications in place, I will show next that sociologists, although without using the philosophical terminology, have engaged in vivid discussions about the merits and limitations of the horizontal and vertical accounts of mechanisms since the 1970s.[6]

1.3.1 Vertical *versus* Horizontal View

Among quantitatively oriented scholars, the confrontation between different approaches to mechanisms already appeared in the muscular critique Hauser (1976) addressed to Boudon's (1974) study of the temporal link between inequality of educational opportunities and social mobility in Western countries. Hauser's most general point was that Boudon did not make use of the best-developed framework for multivariate causal modeling at the time, namely path analysis. According to Hauser, the results that Boudon produced through numerical simulations actually depended on fragile and poorly validated assumptions. Although Boudon (1976) acknowledged that some of Hauser's methodological objections were appropriate, he essentially retorted to Hauser that he missed Boudon's main goal, which was to explore alternative research avenues "to go beyond the statistical relationships to explore the generative mechanisms responsible for them" (*ibid.*: 1187). Boudon's alternative consisted in "ideal-typical models" detailing how the aggregate patterns of interest—which were only summarized, but not explained, by statistical estimates, argued Boudon (1976: 1176, 1178–9, 1183)—can emerge from the dynamic and probably nonlinear relation among the actors' choices and their reactions to other actors' choices, as well as structural constraints (*ibid.*: 1180, 1185–6). A vertical

[5] It should be clarified that this does not mean that horizontal accounts of mechanisms do not, or cannot, consider different levels of analysis. Qualitatively similar information on a certain type of units of analysis can obviously be used to sort these units on various criteria so that clusters of entities are conceptually and numerically represented. In particular, through this procedure, different levels of analysis are represented under the form of different levels of aggregation. This is the typical *modus operandi* within multi-level models for nested data, for instance (see Gelman and Hill 2007). The point rather is the absence within a horizontal perspective on mechanisms of a specific and explicit substantive model in terms of activities and interactions explaining how transitions across levels operate.

[6] On the choice of the terminology "horizontal" and "vertical" accounts of mechanisms, let me add a further qualification. These two adjectives appear sometimes in the critical realism literature to identify a similar distinction as mine between an understanding of social causation based on associations between events and an understanding of causation where these associations are tied to activities and interactions of various entities at various levels of analysis among critical realists (see, for instance, Archer 1998: 196–7). However, this semantic similarity hides an important difference between my view and the critical realist perspective. To me, the notion of level is intrinsically analytical. That is why I constantly adopt the term "levels of analysis" or "levels of abstraction". Thus, the fact that I speak of a "vertical" view on mechanisms should not be taken as a sign of my commitment to a view of social reality as ontologically stratified, a view that I explicitly reject (on the difference between critical and analytical realism, see Di Iorio and Léon-Medina 2021).

view of mechanisms was clearly, but implicitly, at work here. Interestingly, as made more evident by a later article, Boudon (1979) regarded numerical simulations as a necessary tool for this alternative mechanism-based research strategy, although the type of simulations he employed were not, technically speaking, ABMs (on this point, see Manzo 2014b: 435–7).

Recent sociological scholarship shows that the bone of contention is still the opposition between the vertical and horizontal view of mechanisms that was behind the Hauser–Boudon debate. In one of the first meta-theoretical discussions on how the concept of mechanism may reorient empirical research in sociology, Pawson (1989: 130–1) noted that, although a mechanistic representation may have the cognitive function of making a connection between quantitative variables intelligible, it should not be conceptually equated with, nor methodologically operationalized as, a statistical control and/or a set of intervening variables.

This view animated the well-known volume on social mechanisms edited by Hedström and Swedberg (1998), which launched a new wave of discussions on mechanism-based explanations in sociology. As correctly noted by Mahoney (2001: 578, italics added), this new wave of mechanism-based thinking was explicitly motivated by the ambition to go beyond correlation analysis and by the rejection of the view that a mechanism can be simply understood "as an *intervening variable* or set of *intervening variables* that explain why a correlation exists between an independent and dependent variable". Hedström and Swedberg (1998: 17) indeed went back to the Hauser–Boudon debate, attacked the path-analytical tradition in sociology, and ultimately subscribed to the claim that "sociologists in the multivariate modeling tradition still make only rhetorical use of the language of mechanisms". Hedström's (2005: ch. 5) manifesto for analytical sociology endorsed a similar view. Hedström and Ylikoski's (2010: 51–2) review of causal mechanisms in the social sciences also acknowledged the variety of existing accounts of mechanisms but clearly expressed skepticism against the view that mechanisms consist of networks of intervening variables. This positioning is clear when Hedström and Ylikoski (2010: 54) explained that Woodward's (2002) counterfactual account of mechanisms is insufficient because, in fact, "[a] mechanism tells us why the counterfactual dependency [between cause and effect] holds and ties the relata of the counterfactual to the knowledge about entities and relations underlying it", a comment that nicely shows the opposition between Hedström and Ylikoski's endorsement of the vertical view in contrast to horizontal views in the spirt of Woodward.

Kalter and Kroneberg (2014: 101, italics added) echoed this opposition when they noted that, in much quantitative empirical research published in leading European sociological journals, "mechanisms as intervening variables" are "*mistakenly* seen to 'explain' the presumed causal effect of an independent variable on a dependent one". Kalter and Kroneberg lucidly recognized that "mediation analysis"—a crucial statistical tool among those who are animated by a horizontal understanding of mechanisms (for a recent brilliant presentation, see Makovi and Winship 2021)—can be used to test, at least partially, mechanism-based explanations but this requires, they argued, that the postulated mechanism has been previously designed theoretically in a clear and explicit way. In a word, Kalter and Kroneberg concluded, "Mechanisms should not be confused with potential indicators for potential concepts within potential mechanisms."

1.3.2 Horizontal *versus* Vertical View

Within the "horizontal" camp, too, skepticism was overtly expressed against this vertical account of mechanisms as hierarchical dynamic systems of entities, activities, and interactions. After all, as Morgan and Winship (2015: 330) remind us, "[f]or decades, social scientists have considered the explication of mechanisms through the introduction of intervening and mediating variables to be essential to sound explanatory practice in causal analysis". The counterfactual approach to causality—a typical case of the dependence account of causality, in the previous section's terminology—is now reshaping this methodological tradition. As a consequence, the concept of mechanism as a network of intervening variables is also being reshaped in these terms. Kaidesoja (2021b), for instance, explicitly distinguishes between a classical understanding of mechanisms as intervening variables, which he associates to the concept of causality as "robust dependence" (i.e. based on controlling for confounders), and a new version of this understanding of mechanisms as intervening variables where mechanisms are rather seen as sets of counterfactual dependences between variables, which Kaidesoja associated with Woodward's counterfactual interpretations of structural equation models.

This perspective change is especially visible in Knight and Winship (2013: 278). They regard the existing definitions of mechanisms from a vertical viewpoint as "unsatisfactorily vague" and propose a definition that, in their view, better clarifies in what sense a mechanism has a causal structure. According to this view, mechanisms are "modular sets of entities connected by relations of counterfactual dependence" (*ibid*.: 283). The modularity requirement, and the associated condition of "invariance" verified through "ideal" interventions (see Woodward 2003: 98), clearly refer to Woodward's (2002) manipulationist understanding of mechanisms that I discussed above as a typical case of horizontal accounts of mechanisms. In operational terms, however, Knight and Winship (2013: 282, emphasis added) continue to view a mechanism as "(...) a causal relationship involving one or more *intervening variables* between a treatment and an outcome", and they ultimately propose directed acyclic graphs as a framework for discussing under what conditions a net of "mechanistic variables"—the term appears in Morgan and Winship (2015: 335)—allows one to identify causal effects.

From the horizontal viewpoint, dissatisfaction with the vertical view of mechanisms is not only conceptual but also methodological. In this respect, two slightly different, although related, objections may be found in the literature. The first objection is that it is unclear what it means to empirically evaluate alternative hypotheses on mechanisms when mechanisms are regarded as dynamic complex systems of interacting entities bridging levels of analysis. Morgan and Winship (2015: 345) are explicit on this point when they object that the mechanism movement—they refer here to Goldthorpe's (2001) and Hedström's (2005) proposals—runs the risk of falling prey to a "mechanism anarchy", i.e. a proliferation of mechanistic models, with no clear-cut proofs of their empirical significance, or, alternatively, a "mechanism warlordism", i.e. a proliferation of mechanistic models mainly supported by the scientific reputation of their proposers. As a remedy, they suggest a division of labor according to which the "generative mechanism movement", in their own words, contributes to causal inference by developing "how-possible" and "how-plausible" models, while "causal analysis", meaning quantitative techniques for observational data from within a potential outcome approach, provides the tools for assessing the

claims implied by models, which have the pretension to describe actual mechanisms (*ibid*.: 346–7). This proposal is further elaborated by Makovi and Winship (2021) who emphasize now the importance of mediation analysis as a tool to assess rigorously "what portion a variable's effect on an outcome is explained by one or more mediating variables", and reshape the meaning of a mechanism as the set of variables that "mediate some portion of the effect of T on Y". In this sense, they argue, a "mediating mechanism" (M), their own term, "unpacks the black-box of a treatment to outcome relationship by elaborating on how the causal effect is brought about (via M)".

These methodological proposals are in fact motivated by a second objection, which those who regard mechanisms as chains of intervening variables raise against those who regard them as dynamic complex systems of interacting entities bridging levels of analysis. This objection concerns the supposed lack of reliability of computer simulations as a method for building arguments for the existence of the postulated mechanisms (an objection that I will discuss carefully later, see Chapter 4, Section 4.3, and Chapter 5, Section 5.4). Morgan and Winship (2015: 341, fn. 15) formulated this skepticism overtly when they claimed that simulation seems at best a tool for "theory construction", and, even to this task, the tool is of limited utility because of its alleged lack of transparency. One may note that, although this criticism is not supported by a careful discussion of computational simulations, the mistrust of this approach within the "horizontal" camp is not new. By commenting on Hedström and Swedberg's (1998) early volume on social mechanisms, Morgan (2005: 26) already objected that "Sorensen and others got it only partly right. Without a doubt, they correctly identified a major problem with quantitatively oriented sociology. But, they did not offer a sufficiently complete remedy." In short, so the objection goes, no matter how appealing may seem the view of mechanisms as dynamic complex systems of interacting entities bridging levels of analysis, this is a good idea without a sound methodology.

1.4 Causality and Mechanism Accounts, and ABM's Perception

The analysis of the variety of accounts that the concepts of causality and mechanism have received clearly suggests that the different intuitions on the two notions connect in a systematic way. Scholars who have a dependence intuition about causality tend to see mechanisms as chains of intervening variables (horizontal view). In contrast, those who have a production intuition about causality tend to understand mechanisms as complex systems of interacting lower-level units that trigger higher-level outcomes (vertical view). This association is consequential because different views on causality and mechanisms tend to correspond in turn to different ways to open a "black box" underpinning a cause–effect connection, which ultimately lead to different appreciation of the potential contribution of ABMs for causal inference.

In particular, on the one hand, dependence accounts of causality and horizontal views on mechanisms consider that opening a black box amounts to uncovering intermediate variables between a treatment and an outcome. Within this perspective on causality and mechanisms, one should rely on quantitative tools that prioritize finding non-spurious

relationships, establishing counterfactual claims, and, when possible, estimating unbiased parameters that quantify such relations, in a way that allows for the extrapolation from a sample, or test population, to an unobserved target population. As shown by the philosopher Peter Menzies (2012), the horizontal view is indeed typical of the literature on structural equation models and causal graphs (see Pearl 2009). From this perspective, ABM may seem unnecessary to establish causation: what matters is data quality and how creatively one is able to describe these data.

On the other hand, production accounts of causality and vertical views on mechanisms consider that opening a black box means to break the system down into parts and show that the dynamic of the interactions between them can generate, in the sense of reconstruct, the aggregate behavior under scrutiny. Within this perspective, one should rely on methods that prioritize finding a credible narrative that accounts for the observed patterns. The idea is that dependence relations are not constitutive of causality but rather the manifestation of it. From this point of view, simulation methods, and ABMs in particular, would thus seem powerful tools for studying the details of these complex dynamics behind observed connections between happenings. ABM appears as a crucial tool to establish causation (in the sense of production accounts of causality): it provides a formal device to prove that the dependence relationship under scrutiny is deducible by unfolding the postulated (formalized) narrative (see Anzola 2020: 55).

The different reactions to Lucas's (1976) critique to causal inference in macroeconomics (concerning the claim that inflation causes employment) provide a nice historical illustration of the difference between the two camps. In the "dependence" camp, there was a data-driven reaction, which emphasized the centrality of intervention-like methods and led to a more sophisticated use of statistics, the diffusion of time-series econometric models, and the development of vector autoregression methods (Sims 1980), in the tradition of Granger (1969). In contrast, in the "production" camp, there was a theory-driven reaction, which demanded that macroeconomic models be enriched with "micro-foundations". This led initially to the intense use of rational choice models calculating economic aggregates based on individual preferences and expectations—a development encouraged by Lucas (1976) himself—and, consequently, to the critique of representative–agent assumptions (see Kirman 1992; Hoover 2008a, b), to introducing agent-based computational models for solving the aggregation problem in the presence of actors' heterogeneity and various types of social networks (Tesfatsion 2002, 2006; Arthur 2006, 2021). Sociology, and analytical sociology in particular, followed the same path by moving from regression-like statistical methods for survey data to agent-based computational models in order to address the micro-to-macro problem when interdependence structures are present (see Manzo 2020: 200–4).

Some readers may find this opposition exaggerated. Certain ways of framing relationships between different types of causal reasoning may indeed give the impression that the vertical and the horizontal views of mechanisms in fact are closer than it may seem at first glance. This can be seen by looking into the way some scholars have proposed to combine "forward" and "backward" (or "reverse") causal inference (on this distinction, see Section 1.2).

In particular, Sampson et al. (2013) argued that, once the effect of a given cause has been established, we are left with the question of "why" the treatment generated the observed outcome. Thus we need what they call "causal interpretation". To perform this

task, Sampson and colleagues suggest, we have to go back to "backward causation" and design the possible set of causes to which the effect of the specific cause that we have documented belongs. Sampson et al. (2013: 7–8) go until using the term "mechanistic causality" (as a complement of counterfactual causation) to refer to this operation of "causal interpretation" where one tries to unpack the black box underlying the "first-order" effect-of-a-cause dependence documented through a manipulationist approach. This is especially important, Sampson et al. argue, when the goal is to translate the effect of a treatment into policy interventions because, for this goal, we must know why the treatment produces the effect, how this dependence depends on context, and how the effect of the treatment varies across subgroups and over time. Gelman and Imbens (2013) developed a similar line of reasoning when they suggested that "why" questions associated with a causes-of-an-effect approach should be combined with "what-if" questions associated with an effect-of-a-cause perspective. Asking "why" questions, they argue, allows explanatory hypotheses to be formulated for understanding a given documented effect-of-a-cause link as well as to discover errors in the model specification adopted to identify the supposed causal link.

This emphasis on "why" questions may thus seem equivalent to the "why" questions that motivated the "new" vertical view of mechanisms in the philosophy of science and in analytical sociology, which also wants to understand "how" a given connection of interest is brought about. However, the qualitative similarity in terms of cognitive goals—in the sense that, in both cases, one is driven by a need for psychological understanding of the observed connection—translates in different methodological solutions about how the "causal interpretation" should be operated. All examples given by Sampson et al. (2013: figs 2–10) actually amount to "disaggregating" (their own word) the effect-of-a-cause into a causal graph of potentially mediating variables; similarly, Gelman and Imbens (2013) proposed to study the potentially explanatory net of causes-of-the-effect through the same potential outcome approach that was adopted to document the first-order effect-of-a-cause dependence. The possibility of building, and simulating, a formal model of a dynamic system composed of activities, entities, and interactions that could be exploited to show the conditions under which the causal link of interest can appear is not even mentioned.

As the reader knows, no matter how clashing these views on causality, mechanisms, and legitimate methods for causal inference may appear at first, this essay is motivated by the conviction that these views in fact can, and should, be reconciled. Later on, especially in Chapters 4 and 5, I will accumulate elements that suggest that proper causal inference requires a combination of dependence (horizontal) and production (vertical) accounts of causation (mechanisms), thus a synergy between experimental and statistical methods for observational data on the one hand and ABM on the other hand. I will fully develop this argument when defending the "evidential pluralism" thesis (see Chapter 6).

For now, however, let us take at face value the contrasts observed in the literature on causality and mechanisms, and try to explain why, among mathematical models and simulation methods, ABMs can be seen as having a special value for modeling mechanisms from a production point of view on causality and a vertical perspective on mechanism.