4.1 Introduction

4.1.1 Actions, Acts, and Probabilities

Actions admit two interpretations: reactive and deliberative. The reactive interpretation sees action as a consequence of an agent's beliefs, disposition, and environmental inputs, as in "Adam ate the apple because Eve handed it to him." The deliberative interpretation sees action as an option of choice in contemplated decision making, usually involving comparison of consequences, as in "Adam was wondering what God would do if he ate the apple." We shall distinguish the two views by calling the first "act" and the second "action." An act is viewed from the outside, an action from the inside. Therefore, an act can be predicted and can serve as evidence for the actor's stimuli and motivations (provided the actor is part of our model). Actions, in contrast, can neither be predicted nor provide evidence since (by definition) they are pending deliberation and turn into *acts* once executed.

The confusion between actions and acts has led to Newcomb's paradox (Novick 1969) and other oddities in the so-called evidential decision theory, which encourages decision makers to take into consideration the evidence that an action would provide, if enacted. This bizarre theory seems to have loomed from Jeffrey's influential book *The Logic of Decision* (Jeffrey 1965), in which actions are treated as ordinary events (rather than interventions) and, accordingly, the effects of actions are obtained through conditionalization rather than through a mechanism-modifying operation like $do(x)$. (See Stalnaker 1972; Gibbard and Harper 1976; Skyrms 1980; Meek and Glymour 1994; Hitchcock 1996]).

Traditional decision theory [1] instructs rational agents to choose the

[1] I purposely avoid the common title "causal decision theory" in order to suppress even the slightest hint that any alternative, non-causal theory can be used to guide decisions.

option x that maximizes expected utility,[2]

$$U(x) = \sum_y P(y|do(x))u(y)$$

where $u(y)$ is the utility of outcome y; in contrast, "evidential decision" theory calls for maximizing the conditional expectation

$$U_{\text{ev}}(x) = \sum_y P(y|x)u(y),$$

in which x is (improperly) treated as an observed proposition.

The paradoxes that emerge from this fallacy are obvious: patients should avoid going to the doctor "to reduce the probability that one is seriously ill" (Skyrms 1980, p. 130); workers should never hurry to work, to reduce the probability of having overslept; students should not prepare for exams, lest this would prove them behind in their studies; and so on. In short, all remedial actions should be banished lest they increase the probability that a remedy is indeed needed.

The oddity in this kind of logic stems from treating actions as acts that are governed by past associations instead of as objects of free choice, as dictated by the semantics of the $do(x)$ operator. This "evidential" decision theory preaches that one should never ignore genuine statistical evidence (in our case, the evidence that an act normally provides regarding whether the act is needed), but decision theory proper reminds us that actions—by their very definition—render such evidence irrelevant to the decision at hand, for actions *change* the probabilities that acts normally obey.[3]

The moral of this story can be summarized in the following mnemonic rhymes:

---

[2]Following a suggestion of Stalnaker (1972), Gibbard and Harper (1976) used $P(x \mathbin{\Box\!\!\rightarrow} y)$ in $U(x)$, rather than $P(y|do(x))$, where $x \mathbin{\Box\!\!\rightarrow} y$ stands for the subjunctive conditional "y if it were x". The semantics of the two operators are closely related (see Section 7.4), but the equation-removal interpretation of the $do(x)$ operator is less ambiguous and clearly suppresses inference from effect to cause.

[3]Such evidence is rendered irrelevant within the actor's own probability space; in multiagent decision situations, however, each agent should definitely be cognizant of how other agents might interpret each of his pending "would-be" acts.

> Whatever evidence an act might provide
> On facts that preceded the act,
> Should never be used to help one decide
> On whether to choose that same act.

Evidential decision theory was a passing episode in the philosophical literature, and no philosopher today takes the original version of this theory seriously. Still, some recent attempts have been made to revive interest in Jeffrey's expected utility by replacing $P(y|x)$ with $P(y|x, K)$, where K stands for various background contexts, chosen to suppress spurious associations (as in (3.15)) (Price 1991; Hitchcock 1996). Such attempts echo an overly restrictive empiricist tradition, according to which rational agents live and die by one source of information—statistical associations—and hence expected utilities should admit no other operation but Bayes's conditionalization. This tradition is rapidly giving way to a more accommodating conception: rational agents should act according to theories of actions; naturally, such theories demand action-specific conditionalization, (e.g., $do(x)$, while reserving Bayes's conditionalization for representing passive observations (see Goldszmidt and Pearl 1992; Meek and Glymour 1994; Woodward 1995).

In principle, actions are not part of probability theory, and understandably so: probabilities capture normal relationships in the world, whereas actions represent interventions that perturb those relationships. It is no wonder, then, that actions are treated as foreign entities throughout the literature on probability and statistics; they serve neither as arguments of probability expressions nor as events for conditioning such expressions.

Even in the statistical decision-theoretic literature (e.g. Savage 1954), where actions are the main target of analysis, the symbols given to actions serve merely as indices for distinguishing one probability function from another, not as entities that stand in logical relationships to the variables on which probabilities are defined. Savage (1954, p. 14) defined "act" as a "function attaching a consequence to each state of the world," and he treated a chain of decisions, one leading to other, as a single decision. However, the logic that leads us to infer the consequences of actions and strategies from more elementary

considerations is left out of the formalism. For example, consider the actions: "raise taxes," "lower taxes," and "raise interest rates." The consequences of all three actions must be specified separately, prior to analysis, none can be inferred from the others. As a result, if we are given two probabilities, P_A and P_B, denoting the probabilities prevailing under actions A or B, respectively, there is no way we can deduce from this input the probability $P_{A \wedge B}$ corresponding to the joint action $A \wedge B$ or indeed any Boolean combination of the propositions A and B. This means that, in principle, the impact of all anticipated joint actions would need to be specified in advance—an insurmountable task.

The peculiar status of actions in probability theory can be seen most clearly in comparison to the status of observations. By specifying a probability function $P(s)$ on the possible states of the world, we automatically specify how probabilities should change with every conceivable observation e, since $P(s)$ permits us to compute (by conditioning on e) the posterior probabilities $P(E|e)$ for every pair of events E and e. However, specifying $P(s)$ tells us nothing about how probabilities should change in response to an external action $do(A)$. In general, if an action $do(A)$ is to be described as a function that takes $P(s)$ and transforms it to $P_A(s)$, then $P(s)$ tells us nothing about the nature of $P_A(s)$, even when A is an elementary event for which $P(A)$ is well-defined (e.g., "raise the temperature by 1 degree" or "turn the sprinkler on"). With the exception of the trivial requirement that $P_A(s)$ be zero if s implies $\neg A$, a requirement that applies uniformly to every $P(s)$, probability theory does not tell us how $P_A(s)$ should differ from $P'_A(s)$, where $P'(s)$ is some other preaction probability function. Conditioning on A is clearly inadequate for capturing this transformation, as we have seen in many examples in Chapters 1 and 3 (see e.g. Section 1.3.1), because conditioning represents passive observations in an unchanging world whereas actions change the world.

Drawing analogy to visual perception, we may say that the information contained in $P(s)$ is analogous to a precise description of a three-dimensional object; it is sufficient for predicting how that object will be viewed from any angle outside the object, but it is insufficient for predicting how the object will be viewed if manipulated and squeezed by external forces. Additional information about the physical properties of the object must be supplied for making such predictions. By

analogy, the additional information required for describing the transformation from $P(s)$ to $P_A(s)$ should identify those elements of the world that remain invariant under the action $do(A)$. This extra information is provided by causal knowledge, and the $do(\cdot)$ operator enables us to capture the invariant elements (thus defining $P_A(s)$) by locally modifying the graph or the structural equations. The next section will compare this device to the way actions are handled in standard decision theory.

4.1.2 Actions in Decision Analysis

Instead of introducing new operators into probability calculus, the traditional approach has been to attribute the differences between seeing and doing to differences in the total evidence available. Consider the statements: "the barometer reading was observed to be x" and "the barometer reading was set to level x." The former helps us predict the weather, the latter does not. While the evidence described in the first statement is limited to the reading of the barometer, the second statement also tells us that the barometer was manipulated by some agent, and conditioning on this additional evidence should render the barometer reading irrelevant to predicting the rain.

The practical aspects of this approach amount to embracing the acting agents as variables in the analysis, constructing an augmented distribution function including the decisions of those agents, and inferring the effect of actions by conditioning those decision variables to particular values. Thus, for example, the agent manipulating the barometer might enter the system as a decision variable "squeezing the barometer"; after incorporating this variable into the probability distribution, we could infer the impact of manipulating the barometer simply by conditioning the augmented distribution on the event "the barometer was squeezed by force y and has reached level x."

For this conditioning method to work properly in evaluating the effect of future actions, the manipulating agent must be treated as an ideal experimenter acting out of free will, and the associated decision variables must be treated as exogenous—causally unaffected by other variables in the system. For example, if the augmented probability function encodes the fact that the current owner of the barometer tends to squeeze the barometer each time she feels arthritis pain, we

will be unable to use that function for evaluating the effects of deliberate squeezing of the barometer, even by the same owner. Recalling the difference between acts and actions, whenever we set out to calculate the effect of a pending action, we must ignore all mechanisms that constrained or triggered the execution of that action in the past. Accordingly, the event "The barometer was squeezed" must enter the augmented probability function as independent of all events that occurred prior to the time of manipulation, similar to the way action variable F entered the augmented network in Figure 3.2.

This solution corresponds precisely to the way actions are treated in decision analysis, as depicted in the literature on influence diagrams (IDs) (Howard and Matheson 1981; Shachter 1986; Pearl 1988b, chap. 6). Each decision variable is represented as exogenous variable (a parentless node in the diagram), and its impact on other variables is assessed and encoded in terms of conditional probabilities, similar to the impact of any other parent node in the diagram.[4]

The difficulty with this approach is that we need to anticipate in advance, and represent explicitly, all actions whose effects we might wish to evaluate in the future. This renders the modeling process unduly cumbersome, if not totally unmanageable. In circuit diagnosis, for example, it would be awkward to represent every conceivable act of component replacement (similarly, every conceivable connection to a voltage source, current source, etc.) as a node in the diagram. Instead, the effects of such replacements are implicit in the circuit diagram itself and can be deduced from the diagram, given its causal interpretation. In econometric modeling likewise, it would be awkward to represent every conceivable variant of policy intervention as a new variable in the economic equations. Instead, the effects of such interventions can be deduced from the structural interpretation of those equations, if only we can tie the immediate effects of each policy to the corresponding variables and parameters in the equations. The compound action "raise taxes and lower interest rates," for example, need not be introduced as a new variable in the equations, because the effect of that action can be

[4] The ID literature's insistence on divorcing the links in the ID from any causal interpretation (Howard and Matheson 1981; Howard 1990) is at odds with prevailing practice. The causal interpretation is what allows us to treat decision variables as root nodes, unassociated with all other nodes (except their descendants).

deduced if we have the quantities "taxation level" and "interest rates" already represented as (either exogenous or endogenous) variables in the equations.

The ability to predict the effect of interventions without enumerating those interventions in advance is one of the main advantages we draw from causal modeling and one of the main functions served by the notion of causation. Since the number of actions or action combinations is enormous, they cannot be represented explicitly in the model but rather must be indexed by the propositions that each action enforces directly. Indirect consequences of enforcing those propositions are then inferred from the causal relationships among the variables represented in the model. We will return to this theme in Chapter 7 (Section 7.2.4), where we further explore the invariance assumptions that must be met for this encoding scheme to work.

4.1.3 Actions and Counterfactuals

As an alternative to Bayesian conditioning, philosophers (Lewis 1976; Gardenfors 1988) have studied another probability transformation called "imaging," which was deemed useful in the analysis of subjunctive conditionals and which more adequately represents the transformations associated with actions. Whereas Bayes conditioning of $P(s|e)$ transfers the entire probability mass from states excluded by e to the remaining states (in proportion to their current probabilities, $P(s)$), imaging works differently: each excluded state s transfers its mass individually to a select set of states $S^*(s)$ that are considered to be "closest" to s (see Section 7.4.3). Although providing a more adequate and general framework for actions (Gibbard and Harper 1976), imaging leaves the precise specification of the selection function $S^*(s)$ almost unconstrained. Consequently, the problem of enumerating future actions is replaced by the problem of encoding distances among states in a way that would be both economical and respectful of common understanding of the causal laws that operate in the domain. The second requirement is not trivial, considering that indirect ramifications of actions often result in worlds that are quite dissimilar to the one from which we start (Fine 1975).

The difficulties associated with making the closest-world approach

conform to causal laws will be further elaborated in Chapter 7 (Section 7.4). The structural approach pursued in this book escapes these difficulties by basing the notion of interventions directly on causal mechanisms and by capitalizing on the properties of invariance and autonomy that accompany these mechanisms. This mechanism-modification approach can be viewed as a special instance of the closest-world approach, where the closeness measure is crafted so as to respect the causal mechanisms in the domain; the selection function $S^*(s)$ that ensues is represented in (3.13) (see discussion that follows).

The operationality of this mechanism-modification semantics was demonstrated in Chapter 3 and led to the quantitative predictions of the effects of actions, including actions that were not contemplated during the model's construction. The *do* calculus that emerged (Theorem 3.4.1) extends this prediction facility to cases where some of the variables are unobserved. In Chapter 7 we further use the mechanism-modification interpretation to provide semantics for counterfactual statements, as outlined in Section 1.4.4. In this chapter, we will extend the applications of the *do* calculus in several directions, as outlined in the Preface.