Lines of Thought

Central Concepts in Cognitive Psychology

LANCE J. RIPS

OXFORD
UNIVERSITY PRESS

2011

I

Individuals

There are convincing arguments that . . . the most important sort of glue that unites the successive stages of the same persisting thing is causal glue.

David Lewis, "Zimmerman and the Spinning Sphere"

About half way through *The Lady Eve*, there is a scene in which the hero, Charles Pike, is dressing for dinner with the help of his valet, Muggsy. Charles and Muggsy have just encountered Lady Eve, Countess of Sidwich. (Production note: "The Lady Eve [wears] a silver lame dress which I am baffled to describe . . . She looks gorgeous and she knows it.") Muggsy has noticed a perfect likeness between Eve and an earlier character, Jean, with whom Charles had fallen in love but whom he had thrown over after learning that she is a professional card sharp. Here's the dialog in the dressing scene (Sturges, 1985, pp. 465–468):

MUGGSY: That's the same dame . . . she looks the same, she walks the same and she's tossing you just like she did last time.

CHARLES: She doesn't talk the same.

MUGGSY: Anybody can put on an act . . . .

CHARLES: Weren't her eyes a little closer together?

MUGGSY: They were not . . . they were right where they are . . . on each side of her nose . . . .

CHARLES: They look too much alike.

MUGGSY: You said it. They couldn't be two Janes as . . . .

CHARLES: You don't understand me: They look too much alike to be the same.

MUGGSY: That's what I been telling you, they . . . hunh?

CHARLES: You see, if she came here with her hair dyed yellow and her eyebrows different . . . but she *didn't* dye her hair and she didn't pretend she'd never seen me before which is the *first thing* anybody would do. She *says* I look familiar.

MUGGSY: Why shouldn't you?

CHARLES: Because if I did she wouldn't admit it. . . . If she didn't look so exactly *like* the other girl I might be suspicious, but . . . you see you don't understand psychology. If *you* wanted to pretend to be somebody else you'd glue a muff on your chin and the dog wouldn't even bark at you.

MUGGSY: (Indignantly) You tryin' to tell me this ain't the same rib was on the boat? She even wears the same perfume.

CHARLES: (Vacillating) I don't know.

He picks up his dinner jacket and Muggsy helps him on with it. Charles walks out.

Charles's psychological theory works against him here, since it turns out that Muggsy's straightforward reaction is correct. The very last line of the movie belongs to Muggsy: "Positively, the same dame." But is there anything wrong with Charles's reasoning, aside from the fact that he is up against a brilliant woman con artist whose own psychological theorizing is superior to (and able to anticipate) his own? This chapter tries to make the case that, although Muggsy's intuitive response often delivers the right verdict about the identity of people and other objects, Charles is right that identity is a more complicated matter. Our beliefs about connections that unite the earlier and later stages of an object are a main part of our concepts of these individuals.

1.1 Object Concepts and Object Identity

A popular genre in recent historical writing relates what might have happened to important persons but never did. What if Charles I had died of plague in August 1641 (Rabb, 2001)? What if Pizarro had not found potatoes in Peru (McNeill, 2001)? What if Woodrow Wilson had decided not to make public the Zimmerman telegram (Tally, 2000)? Closer to home, we sometimes contemplate contrary-to-fact scenarios in which we play our own starring roles, perhaps in order to figure out what actions might have averted negative events in the past and might avert them in the future (e.g., Kahneman & Miller, 1986; Roese, 1997).

In these speculative moments, we must be able to follow the individuals in question through circumstances in which they never actually existed. Thinking about the results of Wilson failing to disclose the critical telegram, for example, means imagining Wilson's actions in a situation that never occurred. Thinking about what would have happened if you had decided not to go to college entails tracing your path in a world that is only partly like your own.

In spinning these hypothetical stories, we have to identify people and other objects, not only across time and space, but also across alternative histories. This suggests, at a minimum, that our concepts of individuals have to include information that goes beyond what purely perceptual or attentional mechanisms afford. Although we can imagine seeing Wilson carrying out actions that he never iperformed, the mechanisms responsible for such envisioning can't be purely perceptual, since hypothetical events produce no visual traces. Some alternative scenarios would clearly be impossible for Wilson, even in these contrary-to-fact circumstances. Our concept of Wilson helps us decide how he might have acted, and this concept contributes to the overall plausibility of these imaginary events.

The aim of this book is to examine the nature of concepts that are rich enough to support this type of thinking, and we can begin by concentrating on the individual objects that occur in these beliefs and suppositions. I'll use the term *singular concept* to denote a mental representation of a unique individual, and I'll contrast singular concepts with *general concepts*, which are representations of categories. A representation of the Sears Tower is a singular concept in these terms, but our representation of (the category of) buildings or skyscrapers is a general one. The focus in this chapter is on how singular concepts promote judgments of the identity of objects—on how they determine that an object at one time and situation is the same as an object at another time and situation. I'll return to general concepts in Chapter 4, but for now, the basic form of our question is this:

(1) Given knowledge about a target individual x_0 in some situation S, how do we decide whether this individual continues to exist in another (possibly hypothetical) situation S', and if so, which of the individuals $x_1, x_2, \ldots, x_n$ in S' is the same as x_0?

In this context, asking whether individual x_i is the same as x_o means asking whether x_i is *numerically identical* to x_o. This is the equality relation that holds between each specific thing and itself, $x_i = x_o$. Question (1) probes central facts about singular concepts and identity. By comparing potential answers to (1), we can begin to determine which factors are most crucial to our notion of the identity of things.

In deciding among theories of singular concepts, it's helpful to keep the basic properties of the identity relation in mind. According to most treatments, individual identity is reflexive, symmetric, and transitive (e.g., Mendelson, 1964). That is, for any x_i, x_j, and x_k:

(2) a. $x_i = x_i$ (reflexivity).
b. If $x_i = x_j$ then $x_j = x_i$ (symmetry).
c. If $x_i = x_j$ and $x_j = x_k$ then $x_i = x_k$ (transitivity).

People's judgments of identity may sometimes violate these principles, as I discuss later, but the principles provide a starting point for theory development. In order to specify the identity relation more precisely, we can add a fourth principle that is also widely agreed to characterize numerical identity. This principle, sometimes called *Leibniz's Law*, is that if two objects are identical, then any property true of one is also true of the other. This can be expressed as in (3):

(3) For any property F: If $x_i = x_j$, then Fx_i if and only if Fx_j (Leibniz's Law).

Although the principles in (2) and (3) may seem obvious, we will see shortly that it is not always easy to square them with people's judgments about identity.

In the next section of this chapter, I look at earlier cognitive theories of object identity. I argue that these theories are either not powerful enough to explain singular concepts or they rely on overly strong assumptions about the relation between singular and general concepts. I then outline a new theory based on a notion of causal proximity that overcomes some of the earlier theories' difficulties, and I apply it to studies in which participants have to decide whether possible successor objects are identical to an original object. Finally, the last two sections of the chapter discuss how the theory handles problematic cases of identity and compare the model's advantages and disadvantages to those of earlier approaches.

1.2 Theories of Object Concepts

Around Christmas, many of us get cards and accompanying xeroxed letters from friends whom we haven't seen in years. The letters provide news, usually of vacation trips and children's successes, but occasionally of more important life changes, allowing us to keep track of these friends and update our knowledge of them. Our initial encounters with these friends may have given us a rich stock of perceptual information, and this information may survive as part of our mental representation. But unless snapshots come along with these cards, we have to track our friends using nonperceptual facts. Our surviving images may be radically out of date (Bjork, 1978). Still, the Christmas cards may provide enough nonperceptual, descriptive information to allow us to reidentify these people—to determine who in 2011 is the same individual as Aunt Dahlia, whom we last saw in 1970. At a minimum, higher-level information about identity will come into play when perceptual information is absent. Although we sometimes misrecognize people and other objects we know (e.g., Young, Hay, & Ellis, 1985), nevertheless we're often able to keep track of individuals across a night's sleep, lapses of attention, and other perceptual interruptions. Even preschool children can follow individuals over changes in perceptual properties (e.g., Gutheil & Rosengren, 1996; Hall, Lee, & Bélanger, 2001; Hall, Waxman, Brédart, & Nicolay, 2003; Sorrentino, 2001), and they prefer special objects (e.g., their favorite doll or blanket) to perceptually identical duplicates (Hood & Bloom, 2008). We can therefore meaningfully ask what sort of knowledge is relevant to such abilities.

Before introducing a new theory, I outline three earlier ways of looking at this problem of identity of individuals across time. A first possibility makes use of the similarity between object descriptions. An alternative possibility, directed at the identity of concrete, physical objects, depends on the spatial and temporal pathway that an individual follows. According to this proposal, people decide that an individual at an earlier time is the same as one at a later time if and only if a continuous spatiotemporal path connects them. Finally, people's notions of individual identity may depend on knowledge specific to the category in which it belongs. Perhaps people acquire criteria or rules for tracing identity as they learn what kind of thing an individual is. If so, then decisions about

identity across time may be domain specific—different for members of different basic-level categories, such as lions or icebergs.

All three earlier proposals have plausible elements, and people may well employ them in some settings to decide questions of identity. However, each has shortcomings that make it unlikely to serve as a general theory. This section discusses their relative merits, concentrating on theoretical strengths and weaknesses. The following section revisits these proposals and considers their ability to predict new psychological data.

1.2.1 Similarity

A simple answer to the question of how we determine that items are identical is that we use our knowledge of their common and distinctive properties. If we can compute a measure of the similarity between the items from these properties, we could then judge the items identical if their similarity exceeds some threshold. This is a proposal we should take seriously for several reasons. For one thing, similarity seems to influence perceptual impressions of identity. When observers are shown two line segments, one after the other, rapidly alternating, they are more likely to see a single line segment "moving" than to see two distinct line segments if the segments have a similar orientation (Ullman, 1979; for related results with polygons, see Farrell & Shepard, 1981). For another, recognition of both individual words (e.g., Anisfeld & Knapp, 1968) and pictures (Bower & Glass, 1976) is sensitive to the similarity between the originally presented items and the test items.[1] Similarity almost inevitably plays a role in judging identity. If a cat runs behind a couch and a very similar looking cat runs out the other side, we take this similarity as indicating a single cat in the absence of information to the contrary (such as the presence of twin cats). This is the instinct that Muggsy goes on.

However, a pure similarity theory runs into some difficulties, and these difficulties provide a secondary theme in this book (see Sloman & Rips, 1998, for the status of similarity in cognitive models). First, properties of the items in question are likely to contribute unequally to judgments of identity. Aunt Dahlia's taste in music and other matters in 1970 (mostly Motown) may be vastly different from her taste in 2011 (mostly Mahler), so that her taste and preferences in 1970 and 2011 may differ in ways irrelevant to her identity. We therefore need a theory about which

properties are relevant to judgments of identity—a variation of the question with which we started.

Second, similarity may presuppose identity (as Fodor & Lepore, 1992, argue). If we use properties of Dahlia_{1970} and Dahlia_{2011} to establish the similarity between them, then we must be able to determine that these properties are the same. For example, if near-sightedness is one such property, we need to know that $\text{Dahlia's-nearsightedness}_{1970}$ = $\text{Dahlia's-nearsightedness}_{2011}$. But this shifts the problem from sameness of objects to sameness of properties.

Third, things change. We should expect some of Aunt Dahlia's properties to change in predictable ways over time, and although these changes make for dissimilarities, they should count for, rather than against, the possibility that a later stage belongs to the same individual as her earlier stages (Rosengren, Gelman, Kalish, & McCormick, 1991; Sternberg, 1982). An individual at three years of age would typically be shorter than, not the same size as, the same individual at 23. If the individual x_0 is 3′5″ in 1989 and x_1 is 3′5″ in 2011, that could be evidence they were *not* identical. Along the same lines, people sometimes perceive identity despite radical dissimilarities. In a well-known demonstration by Simons and Levin (1998), an experimenter asked a pedestrian for directions on a university campus. While the pedestrian was engaged in the conversation, confederates barged in front of the pedestrian, carrying a door that momentarily concealed the experimenter. During the concealment, one of the confederates exchanged places with the experimenter, continuing the interaction with the pedestrian. But despite the fact that the two experimenters were not especially similar in appearance, only about half the pedestrian participants noticed the change in identity.

For these reasons, similarity is limited in what it can do as a theory of object identity. Although "identical" twins can be amazingly similar, they can't be truly identical.

1.2.2 Spatiotemporal Continuity

According to the continuity view, we judge two individuals identical if we know that these individuals fall on the same unbroken spatiotemporal path. Dahlia_{1970} = Dahlia_{2011}, for example, if we can find a continuous path linking the first to the second. This theory is similar to one that

is sometimes offered for perceptual tracking (e.g., Spelke, Kestenbaum, Simons, & Wein, 1995). There is also evidence (Stone, 1998) that the spatiotemporal path that an object takes can influence later recognition of that object: Recognition is better if observers see the same path at test than if they see an equally informative alternative path.

Although the continuity theory is more substantial than the similarity proposal, counterexamples suggest it won't work. D. M. Armstrong (1980), Nozick (1981, pp. 655–656), and Shoemaker (1979) provide a thought experiment of this sort:

> *Dual-ing machines*: Imagine two machines: one capable of vaporizing an object and the other capable of materializing an object in an arbitrarily brief interval. Suppose, too, that these machines operate on completely independent schedules so there is no connection between one machine and the other. It is possible to conceive the first machine vaporizing a specific object—say, a chair—and the second machine, by chance, immediately materializing a qualitatively similar but distinct object without a temporal gap and in exactly the same spatial location. Under these circumstances, an observer would notice no change whatever, since nothing about their spatial or temporal position or their qualitative properties would distinguish the vaporized and materialized chairs from a single chair. But in the imagined scenario, although an unbroken spatiotemporal sequence of chair stages exists, there are two chairs in play rather than one.

If this example is correct, people may always be willing to override purely spatiotemporal information if they know enough about the facts of the case. For any imagined spatiotemporal evidence pointing to one object, we can conceive Armstrong-Nozick-Shoemaker machines that substitute multiple objects. The example also shows that any sort of perceptual evidence for identity will also be insufficient.

One reaction to this example is that the machines don't truly preserve spatiotemporal continuity; there must be some break between the two chairs, since they have different material composition or other properties. But although a difference exists between the chairs, it needn't be a *spatial* or *temporal* discontinuity. We can envision the materializing machine outputting the second chair within any temporal interval ε ($\varepsilon > 0$)

following the disappearance of the first chair. If so, this meets the standard definition of continuity. Of course, this example depends on contemplating sci-fi devices that may never actually exist, but the fact that we can make sense of dual-ing machines suggests that we don't conceive of spatiotemporal continuity as guaranteeing identity over time.

Hirsch (1982) provides a second type of counterexample to the idea that spatiotemporal continuity is sufficient for identity. As Hirsch points out, indefinitely many spatially and temporally continuous sequences don't count as a single object. The north half of a cat from 10 to 11 pm is one such nonobject.

To be sure, the dual-ing machine example does not show that all forms of continuity are irrelevant for identity. Intuitively, the reason the example describes two chairs rather than one is that the vaporized chair is not connected to the materialized one in the same way as the successive stages of a single chair. In particular, there is no causal link between the vaporized and the materialized chairs. This intuition leads directly to the theory of identity that I propose later. Incorporating causal relations takes us a significant step beyond spatial and temporal continuity.

It also seems doubtful that spatiotemporal continuity is necessary for identity. We could disassemble a computer into its individual circuit components, store the resulting hundreds or thousands of parts in separate locations, and then reassemble the parts later in yet another location but according to exactly the same pattern. Under these circumstances, the later reconstructed computer would seem to be identical to the earlier intact one. However, no continuous spatiotemporal path links the two halves of the computer's existence. This implies that identity is possible over gaps in time and space (as Hirsch, 1982, argues from a similar example).[2]

The computer example should make us cautious about requiring continuity as a criterion of identity, but the example also hints at another basis for singular concepts. Computers, tables, chairs, cars, and many other artifacts can survive complete disassembly and reassembly, but cats, robins, roses, and many other living things can't survive total dismemberment. Some evidence that older children and adults recognize such a distinction comes from Hall (1998), which I discuss in Section 1.4. Perhaps, then, identity over time is relative to the category to which an object belongs. A common theme in cognitive research on categorization is that knowledge of an object's category can provide theoretical information

about the object, information that fuels inference and prediction (see Chapters 4 and 5, in this volume). The theory I'm about to take up extends this idea by supposing that category-level concepts also supply criteria for identifying category members from one moment to the next.

1.2.3 Sortals

Certain concepts may determine rules for individuating and identifying their instances. The concept of cats, for example, may consist in part of rules for differentiating individual cats in a mass of cats-and-other-objects and identifying each cat over time. (Not all psychological theories assume that general concepts contain rules, as we will see in Chapters 4 and 5, but the theory we're about to discuss presupposes them.)

Philosophical work discusses this idea under the heading of *sortals* (Strawson, 1959, p. 168). A sortal is a count noun like *table* that is capable of singling out individual tables. By contrast, an adjective like *black* denotes a property that doesn't by itself distinguish things. We can't count the black stuff that composes a black table, say, since the total is indeterminate: It might be one thing (the table), five (the legs + the top), six (the legs + the top + the table), or more. Nouns like *table*, *leg*, or *top*, however, do provide the resources we need to get a determinate answer. Orthodox sortal theories assert that there are no individuals at all, apart from the sortal concepts that carve them out and establish their beginnings and endings. As Dummett (1973, p. 179) puts it, "[John Stuart] Mill wrote as though the world already came to us sliced up into objects, and all we have to learn is which label to tie on to which object. But it is not so: the proper names which we use, and the corresponding sortal terms, determine principles whereby the slicing up is to be effected, principles which are acquired with the acquisition of the uses of these words." In what follows, I'll use the term *sortal* for linguistic expressions (i.e., for certain count nouns), *sortal concept* to refer to the associated mental representation, and *sortal category* for the referent of the sortal (a set of objects).[3]

The identity conditions that these sortals furnish are necessary and sufficient relations for identifying objects at different times, and they take the form in (4) (see Lowe, 1989):

(4) If object x at time t_1 and object y at time t_2 are members of sortal category S, then $x = y$ if and only if $R_S(x,y)$.

In this formulation, R_S is an equivalence relation (i.e., it is reflexive, symmetric, and transitive; see (2) above). In addition, R_S must be an informative relation—one that does not merely paraphrase or presuppose identity for *S*'s—in order to avoid trivializing the analysis. The plausibility of sortal theory will, of course, depend on whether a relation exists that can fill R_S's role in (4). In the case of formal or mathematical categories, the relation is sometimes obvious. For example, we can define the identity of two sets in terms of the same-member relation: Sets x and y are identical if and only if x has as its members all and only the members of y. In the case of natural categories, however, the existence of an appropriate R_S might be in doubt, an issue we will return to in Section 1.5.2. Most sortal theories also assume that objects in distinct sortals cannot be identical. In other words:

(5) If object x at time t_1 is a member of sortal category S and object y at time t_2 is not a member of S, then $x \neq y$.[4]

Some psychologists have enlisted sortals to explain how people trace the history of individuals. An individual object, such as a cat or a table, can undergo a variety of changes without ceasing to exist, whereas other changes are not compatible with its continued existence. By (5), the compatible changes can't take an individual outside its sortal category. An individual x_0 can't persist as an individual x_1 if x_0 is in sortal category S_0 and x_1 is in a different sortal category S_1. Which changes are possible and which impossible vary across types of objects: They are relative to an object's sortal. Some changes—such as total disassembly—may be possible for a table but not for a cat. According to Wiggins (2001), an object's sortal (e.g., *table* or *cat*) is the term that best provides the answer to the question "What is it?" for that object.

One advantage of the sortal theory is that it handles some issues that are problematic for continuity accounts. Consider, for example, a car that loses a hubcap on a bumpy road. Although both the hubcapless car and the carless hubcap are continuous with the original item, the sortal *car* applies to the initial object and dictates that it's the hubcapless car that is identical to the original car. A second advantage is that the sortal theory can help explain the problem with which we started: how we are able to make judgments about object identity even in counterfactual contexts. Because sortals separate possible from impossible changes for the objects they apply to, they rule out some alternative histories for

an object. These theoretical advantages suggest that it might be useful to incorporate the sortalist insights in psychological explanations of object identity, so we need to examine carefully some recent attempts of this kind.

1.2.3.1 Sortalist Approaches in Psychological Theories

For researchers who see deficiencies in pure similarity and continuity accounts, sortals fill a gap by providing a source of rules that people can use to keep track of things. In this vein, Macnamara (1986) assumed on theoretical grounds that when children learn a proper name for an object, they interpret the name with the help of the object's sortal concept. The sortal concept—which Macnamara took to be a prototype or perceptual "gestalt" of a category—provides criteria for individuation and identity that support correct use of the proper name. The same considerations apply to the use of personal pronouns, such as *I* and *you* (Oshima-Takane, 1999). Similarly, according to Carey (1995a, p. 108), "To see the logical role sortals play in our thought, first consider that we cannot simply count what is in this room. Before we begin counting, we must be told what to count. We must supply a sortal . . . Next consider whether a given entity is the same one as we saw before. Again, we must be supplied a sortal to trace identity."

Xu and Carey (1996; Xu, Carey, & Quint, 2004) recruit sortals to explain results on the way infants discriminate objects. In these experiments, infants viewed an opaque screen from which objects emerged, either at the right or left. An infant might see, for example, a toy elephant emerge from the right side of the screen and then return behind the screen. A short time later, a cup emerges from the left side of the screen and returns behind the screen. This performance is repeated a number of times with the same two objects. The screen is then removed, revealing either a single object (e.g., the cup) or two objects (cup and elephant). The data from these experiments show that, when the screen is removed, 10-month-old infants look no longer at the scene with one object than at the scene with two (relative to baseline performance). By contrast, 12-month-olds look longer at the one-object tableau, as long as the objects are from different basic-level categories. (If the objects are from the same category—e.g., two cups with different colors—even

12-month-olds fail to look longer at the one-object scenes; see Xu et al., 2004.) Xu and Carey interpret this to mean that the younger infants do not expect to see two objects and so are no more surprised by one than by two in this context. Older infants, however, can use their knowledge of the sortal concepts CUP and ELEPHANT to make the discrimination.[5]

Several factors allow younger infants to anticipate two objects correctly. First, if the 10-month-olds are able to inspect simultaneously the elephant and the cup before the start of the trial, then they do stare longer at the one-object scene (Xu & Carey, 1996). Second, if the experimenter labels the two objects differently ("look a blicket" vs. "look a gax") while they are moving back and forth, 10-month-olds again perform correctly (Xu, 2002). This combination of results suggests, according to Xu and Carey, that younger infants can use spatial or verbal cues to individuate the objects; without these cues, they are unable to anticipate the presence of two objects, since they don't know that elephants don't morph into cups while briefly out of sight.

According to Xu and Carey (1996), the younger infants who fail this is-it-one-or-two task lack knowledge of sortal concepts (e.g., CUP, [toy] ELEPHANT) that would allow them to individuate the objects conceptually. Since this individuating information is supposed to be a crucial part of the meaning of sortals, these infants don't know these meanings; they don't have adult-like concepts for even basic-level categories such as *cup*.[6]

1.2.3.2 Evidence Concerning Sortals

Carey and Xu (2001; Xu, 2003a) maintain that infants acquire the meaning of these sortals at about 12 months of age and that the sortals are responsible for older infants' and adults' correct performance in the is-it-one-or-two task.[7] Is there psychological evidence that *cup*, *elephant*, and similar count nouns play this identifying role?

One way to investigate this issue takes advantage of Wiggins's (1997, 2001) contention that the sortal for a particular object is the term that answers the question, "What is it?" Brown (1958) and Rosch, Mervis, Gray, Johnson, and Boyes-Braem (1976) have claimed that words for basic-level categories, such as *cup* or *elephant*, usually provide the answer

to this question; so we may be able to check whether knowledge of basic-level categories gives people the means to identify objects.[8] In one attempt of this kind, Liittschwager (1995) gave 4-year-old children illustrated stories about people who were magically transformed to different states. The transformations ranged, across trials, from simple within-category changes in properties (e.g., from a clean to a dirty child) to more extreme cross-category changes (e.g., from a girl to a cat or from a woman to rain). For each type of transformation, participants decided whether the transformed object could still be called by the name of the original person—for example, "Do you think that now *this* is Ali?" According to sortal theories, objects cannot maintain their identity across changes in sortal categories (see Principle (5)); so participants should use the same proper name only if the transformation is within the basic-level category person. The results of this study showed that as the transformational distance increased between the original person and the final product, participants were less willing to apply the proper name. However, there was no discernible elbow in this function at the sortal category—the boundary between persons and nonpersons. According to Liittschwager (1995, pp. 33–34), the data "provide little support for Macnamara's (1986) position that proper names should be maintained across changes up to (but not beyond) the basic level."

Sergey Blok, George Newman, and I (Blok, Newman, & Rips, 2005) report a similar finding in an experiment that also employed transformation stories. Participants (college students) read stories about an individual—say, Jim—who has a severe traffic accident in the year 2020 and whose only hope for survival is radical surgery. In the condition most relevant for present purposes, participants learned that Jim's brain was transplanted to a different body. On some trials, scientists placed the brain in "a highly sophisticated cybernetic body," while on others they placed it in a human body that scientists had grown for just such emergencies. In each case, Jim's old body was destroyed. The stories described the operation as successful in allowing the brain to control the new body, but participants also learned either that Jim's memories survived the operation intact or did not survive. After reading the scenario, participants rated on a 0-to-9 scale their agreement with each of two statements: (a) The transplant recipient is Jim after the operation, and (b) the transplant recipient is a person after the operation.[9]

The results from Blok et al. (2005) show a dissociation between identity and category judgments. Figure 1.1 displays the mean agreement ratings as a function of whether the story described the brain transplanted to a robot or to a human body and also whether the memories survived or did not survive the operation. Participants were more likely to agree that the post-op recipient was still Jim (open circles in Figure 1.1) if Jim's memories were preserved. However, there was a much smaller effect of whether these memories were embodied in a human or in a

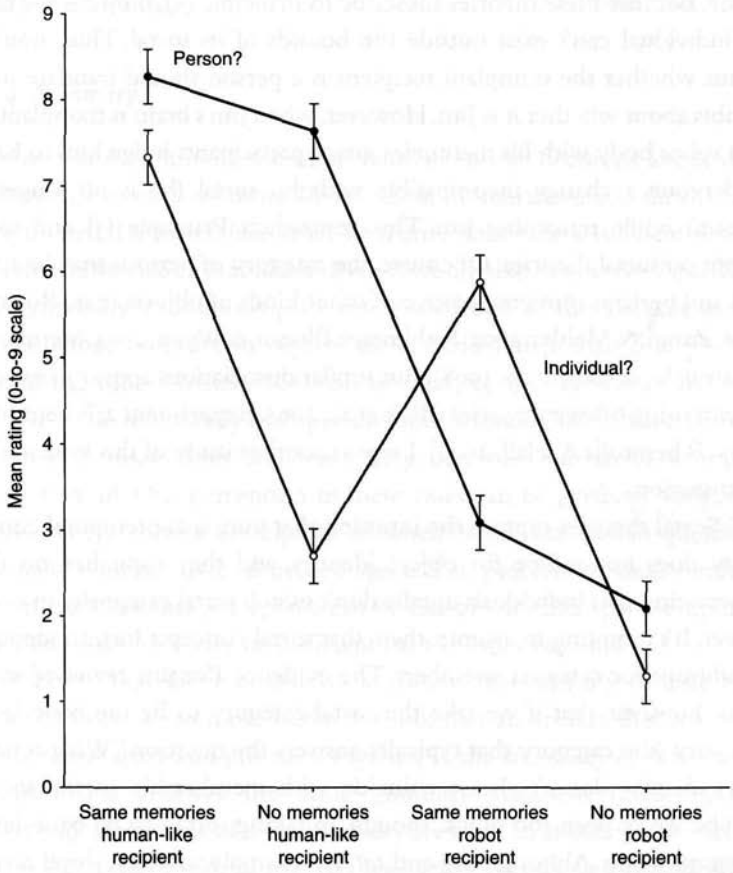

FIGURE 1.1 Mean agreement ratings (0-to-9 scale) for the statements that the transplant recipient is "still Jim" (open circles) and "still a person" (closed circles). The x-axis represents four versions of the accident story. Error bars indicate ±1 standard error of the mean (from Blok et al., 2005).

robot body (see also Nichols & Bruno, 2010). Ratings about whether the recipient was a person (filled circles), however, show the opposite pattern. Participants were more likely to think the transformed object was a person if it had a human rather than a robot body, but they relied less heavily on whether Jim's memories remained intact. This combination of effects produced the finding that when Jim's memories survive in a robotic body, participants were much more likely to think that the transformed individual is Jim than that it (!) is a person.

Psychological versions of sortal theories seem at odds with this outcome. Because these theories subscribe to Principle (5), they assume that an individual can't exist outside the bounds of its sortal. Thus, doubts about whether the transplant recipient is a person should translate into doubts about whether it is Jim. However, when Jim's brain is transplanted to a robot body with his memories intact, participants judge him to have undergone a change incompatible with his sortal (he is no longer a person) while remaining Jim. This contradicts Principle (5) and casts doubt on sortal theories. Of course, the category of persons may be special and perhaps unrepresentative of other kinds of objects (e.g., Bonatti, Frot, Zangl, & Mehler, 2002; Kuhlmeier, Bloom, & Wynn, 2004; Sternberg, Chawarski, & Allbritton, 1998). But similar dissociations appear in experiments using other categories (Blok et al., 2005, Experiment 2; Rhemtulla, 2005; Rhemtulla & Hall, 2009). I report another study of this type in the next section.

Sortal theories capture the intuition that pure spatiotemporal continuity does not suffice for object identity, and they capitalize on the observation that individuals usually don't switch sortal categories in mid-career. It's tempting to assume, then, that sortal concepts furnish identity conditions for category members. The evidence I've just reviewed suggests, however, that if we take the sortal category to be the basic-level category (the category that typically answers the question "What is it?") then identity doesn't always coincide with membership in the sortal. Maybe we've been too quick, though, in taking sortals to be basic-level category terms. Although *dog* and *cat*, for example, are basic-level terms and seem to answer the "What is it?" question, we may establish Rover's or Cat-astrophe's criteria of identity at some other level. Since dogs and cats are similar biologically, their identity may depend on general principles at the level of mammals, vertebrates, or even animals (see Dummett,

1973, p. 76; but also Wiggins, 2001, p. 129, for an explicit denial that *animal* is a sortal). It appears to be a difficult matter to pin down exactly which term is the appropriate sortal, as Mackie (2006) has pointed out. But if we cut the tie to basic-level categories, we seem to be left with a fairly weak hypothesis. A softened version of the sortal theory might claim that for any object there is some count noun or other whose concept provides identity conditions for the object. But all the softened theory does is narrow down the identity conditions to those that a count noun can represent. And this seems little different from the claim that objects have identity conditions.

1.2.4 Summary

Similarity and continuity are often relevant to our decisions about the identity of objects over time. We use them as heuristics, and they often serve us well. However, like other heuristics, they take a backseat when definitive information is available. Even if we are able to observe a spatially and temporally continuous path, our knowledge of the circumstances may convince us that two objects are at hand rather than one (D. M. Armstrong, 1980; Nozick, 1981; Shoemaker, 1979). Likewise, what we perceive to be temporally and spatially discontinuous, like a table before and after its disassembly and reassembly, may turn out to be a single object after all. Our perception in these cases can be perfectly veridical in providing a correct description of whether the material in question is actually continuous in time and space. The problem in determining identity isn't just that perception can be illusory; it's that spatiotemporal continuity isn't necessary or sufficient for identity over time.

People adopt different criteria in judging the identity of different types of things. We're more likely to think that an artifact that is completely taken apart and put back together is the same object than is an organism that receives the same ghoulish treatment (Hall, 1998). Psychological versions of sortal theories are right in stressing this variation in our criteria. Sortal theories also seem correct in asserting that identity over time depends on the kinds of changes that are possible or impossible for an individual. What's more controversial is the source of identity principles. The evidence to date suggests that these principles are not simply inherited from basic-level concepts. Our decisions about

whether items are identical depend on induction over a potentially broader knowledge base.

1.3 A Causal Continuer Theory of Object Identity

This section considers a theory of identity judgments that draws on some of the elements of the earlier views just surveyed, but combines them in a new way—one that is consistent with the perspective on general concepts that I develop later in this book (see Chapter 4). The theory attempts to answer the question posed in (1) by describing the cognitive processes people go through when they have to decide whether an individual object, x_0, existing at one time, is identical to one of a set of candidate objects $x_1, x_2, \ldots, x_n$, existing at a later time. The model derives from a proposal by Nozick (1981)—his Closest Continuer theory— but I recast the proposal here as a descriptive psychological account. I intend the model to help explain how people judge object identity, and it may not necessarily give a correct account of the underlying (metaphysical) nature of identity over time. I first describe the theory in outline and then apply it to some experimental data that Blok, Newman, and I have collected to test a quantitative version of this approach.

1.3.1 The Causal Continuer Theory

As its name implies, Nozick's (1981) Closest Continuer theory commits itself to the idea that the object identical to the original x_0 is the one that is, in some sense, "closest." But unlike the similarity approach, discussed in the preceding section, the present theory determines closeness within a framework of causal principles. As a reminder of this restriction (and some additional modifications, mentioned later) I refer to the model proposed here as the *Causal Continuer* theory.

Causality is important in this context, since the theory's chief principle is that the continuer of the original object must be a causal outgrowth of that original. Here's a story to illustrate this idea:

> *The missing chair:* Suppose you own a chair with a distinctive color and shape. One day you regrettably leave the chair in one of the

> classrooms in the department, from which it disappears. The following week you spot two different chairs that are qualitatively identical to yours. One is sitting in the office of Professor A; the other in the office of Professor B. Which, if either, of these chairs is yours? Similarity is clearly unable to decide the case. Spatiotemporal continuity might be helpful if you could establish that there is a continuous pathway from the chair in the classroom to the chair in one of the offices. But suppose that on investigating you find the case is this: Professor A, who had never seen your chair, happened to construct one of the same shape and color. Professor B, however, has disassembled the chair he found in the classroom, stealthily moved the parts to his office one at a time, and reassembled the chair. No spatially continuous path connects your chair to either chair, but there is a clear intuition that the chair in Professor B's office, and not the one in Professor A's, is yours. A causal relation links each step in the transition from the chair in the classroom to the chair in Professor B's office, but no such causal relation exists between your chair and Professor A's.

The important role that causality plays in the theory goes along with the idea that causal forces are central in producing an object, maintaining it through time, and eventually destroying it. In this respect, the Causal Continuer theory is akin to psychological essentialism (S. A. Gelman, 2003; Medin & Ortony, 1989), which also emphasizes the role of causality in people's thinking about natural kinds (see Chapter 4 in this book for a discussion of essentialism). It also agrees with some versions of psychological essentialism in supposing that separate causal factors are responsible for category membership and individual persistence (Gelman, 2003; Gutheil & Rosengren, 1996). However, the present theory takes no stand on a unique, distinctive cause that would answer to the notion of an essence. The existence of an object may be a function of many conspiring causes, some internal and some external to the object (see Sloman & Malt, 2003; Strevens, 2000; and Chapter 4 of this book). Moreover, the causes governing a category member may partially overlap those governing its category. For example, respiration, circulation, and many other bodily causal systems may be necessary for the survival of both an individual organism and the species to which it belongs.

Similarly, the Causal Continuer theory also makes contact with recent models of categories that emphasize the role of causality in category structure (e.g., Ahn, 1998; Rehder & Hastie, 2001). For reasons that I have mentioned in connection with sortals, however, I assume that the causes responsible for an individual's persistence may include not only those associated with its basic-level category, but also the larger set of background causes that governs the environment in which the individual finds itself. Otherwise, it would be difficult to explain the dissociation between Jim's continued existence as Jim and his continued existence as a person in the experiment reported earlier.

A second aspect of the theory is that, in determining a continuer, we cannot select something that is arbitrarily far from the original. In some later situations, *no* object may qualify as identical to the one with which we started. Although later objects may causally stem from the original, the causal connections to those objects may be so attenuated that none can serve as a continuer, and the original object thereby goes out of existence. If a book is ripped apart into its covers and its individual pages (each page separated from the others), then each of the resulting pieces maintains a causal connection to the original, but the connection may not be strong enough to qualify any of the pieces (or their sum) as the book. Similarly, the causal connection between the original object and a later one cannot be too abrupt. Although the dead remains of an animal are causal products of its living state, the transition is not smooth enough to allow the remains to serve as a continuer of the organism.

We can think of these restrictions as imposing a two-step decision process. To determine which of a set of objects at a later time is identical to an original: (a) we consider only those later objects whose connection to the original exceeds some threshold (no other objects can be continuers), and (b) within the range of close-enough objects, we select the closest as the one identical to the original. It may seem natural to assume that people carry out step (a) before step (b), but the opposite ordering is also possible. People may identify the closest object before determining whether that object is close enough to be identical. Note, too, that step (b) allows the decision process to be context sensitive. An item that is closest in one situation may not be closest in another if the second situation contains an even closer object.[10]

In Nozick's (1981) theory, the closest continuer must be closest in an absolute sense—no ties are allowed. For example, if an amoeba x_0 divides in such a way that the two descendants, x_1 and x_2, are equally like their common parent, then the parent cannot be identical to either descendant. The reason for this additional restriction has to do with the transitivity of identity, which we glimpsed in (2c). The two amoeba descendants, x_1 and x_2, aren't equal to each other, since each can go its own way, acquiring different properties after the division that produced it. But then if the parent x_0 is equal to both the descendants, the result is an intransitivity: $x_1 = x_0$ and $x_0 = x_2$, but $x_1 \neq x_2$. However, similar apparent intransitivities arise in certain perceptual situations (Ullman, 1979), and for this reason I leave room for the possibility of ties in judgments about conceptual identity. If such judgments do exist, we can then consider how to interpret them.[11] I'll take up this issue in more detail in Section 1.5.1.

In examining the theory, I concentrate on the basic two-part decision structure just outlined. The rest of this section presents two experiments that carry out such an examination, and the Appendix formulates a quantitative version of the theory that applies to the data.

1.3.2 An Experiment on Individual Persistence

To find out how well the Causal Continuer theory handles people's identity judgments, we need an experimental situation that gives participants a choice between potential continuers and that varies the causal distance between the continuers and the original object. Because the effect of category membership is of interest (as an additional test of sortal theories), the original object must be able to switch categories. These requirements are difficult or impossible to satisfy with everyday objects, but we can approximate such situations in stories about hypothetical transformations, as in much earlier research on concepts and categories (e.g., Blok et al., 2005; S. A. Gelman & Wellman, 1991; C. N. Johnson, 1990; Keil, 1989; Liittschwager, 1995; Rips, 1989). It's good to keep in mind, however, that the Causal Continuer theory applies to everyday identity decisions, as well as to the recherché cases we consider here. The purpose of using hypothetical scenarios is the usual one of achieving experimental control over variables that are confounded

in typical situations. Experimental control is nearly always in tension with ecological validity. In Section 1.3.4, however, we will look at an experimental setting that may be closer to the usual contexts in which identity is in question.

The stories I used in this experiment are similar to those in some philosophical discussions (e.g., Lewis, 1983; Nozick, 1981; Parfit, 1984; Perry, 1972) and described a "transporter" that could copy and transfer objects from place to place on a particle-by-particle basis. The copied particles are transmitted to a new location and put back together according to a blueprint of the original. The particles of the original are entirely destroyed in the copying process. Thus, there was no spatiotemporal or material continuity between the original and the copy, but the copy causally stems from the original by means of the duplicating process. (This explicit causal relation distinguishes this set up from the one in the dual-ing machines example.) Each trial of the experiment described a different hypothetical transformation, and participants' task was to make two decisions about the resulting copies: (a) whether the copy is the same object as the original, and (b) whether the copy is in the same category as the original.

The experimental stories included three variations. First, they varied whether there was one copy or two. In one block of trials, the instructions told participants that the transporter had made a single copy of the particles, and the participants decided whether that copy was identical to the original and whether it was in the same category as the original. On a second block of trials, the instructions stated that the transporter constructed two copies. Participants then decided whether one, both, or neither of these copies was identical to the original, and whether one, both, or neither was in the original's category. The second variation among the stories concerned the percentage of copied particles that went into the reconstituted objects. In the one-copy condition, the copy could contain 0, 25, 50, 75, or 100% of the particles copied from the original. The story specified that in the 0, 25, 50, or 75% conditions the remaining particles came from a different object. In the two-copy condition, each copy could independently contain any of the five percentages just mentioned, with the residual particles again coming from a different object. For example, participants might learn that one copy included 50% particles coming from the original object and 50% from a separate

object, while the second copy included 75% particles from the original and 25% from the separate object. (The percentage of particles from the original needn't add to 100%, since the transporter was said to have made two complete batches of particles.) In the context of this experiment, the percentage of particles from the original object provides a measure of the causal distance between the original and each of the copies. Finally, the stories also varied whether the residual particles came from a member of the same category as the original or from a member of a different category. In each story, the original item was a lion (named "Fred"), and the residual particles were either from a second lion ("Calvin") or from a tiger ("Joe"). Thus, in the one-copy condition, participants might learn on one trial that the newly constructed creature contained 75% particles from Fred and the remaining 25% from the same-category member, Calvin. On another trial, the creature contained 75% particles from Fred and the remaining 25% from the different-category member, Joe. In the two-copy condition, both copies had residual particles from the second lion or both had residual particles from the tiger.

The instructions explained the workings of the "transporter" in the same way that I described it earlier. In the one-copy condition, the participants (Northwestern University students) received nine scenarios that differed in the percentage of particles coming from the original object and in the source of the remaining particles. (There were nine rather than ten scenarios, since when 100% of particles were from the original, there were no residual particles and thus no possible difference in source.) On each trial, participants made separate decisions about whether the outcome of the transformation was the same individual (they chose between "Is Fred" or "Is not Fred") and whether it was a member of the same category ("Is a lion" or "Is not a lion"). In the two-copy condition, participants received 30 two-copy trials. For each story, they again made an individual decision (they selected one of: "Only Copy A is Fred," "Only Copy B is Fred," "Both copies are Fred," or "Neither copy is Fred") and a category decision ("Only Copy A is a lion," "Only Copy B is a lion," "Both copies are lions," or "Neither copy is a lion").

I focus on the results of the two-copy condition, since they provide the best test of the model, and use the one-copy condition mainly to estimate parameters associated with causal distance between the original object and each alternative (see the Appendix to this chapter). However,

the one-copy data also provide some evidence about which of the experimental factors affect decisions about individual identity and about category membership. The results appear in Figure 1.2, and they exhibit a clear dissociation between these two types of judgments, confirming the conclusions from Blok et al. (2005) discussed earlier. The x-axis in

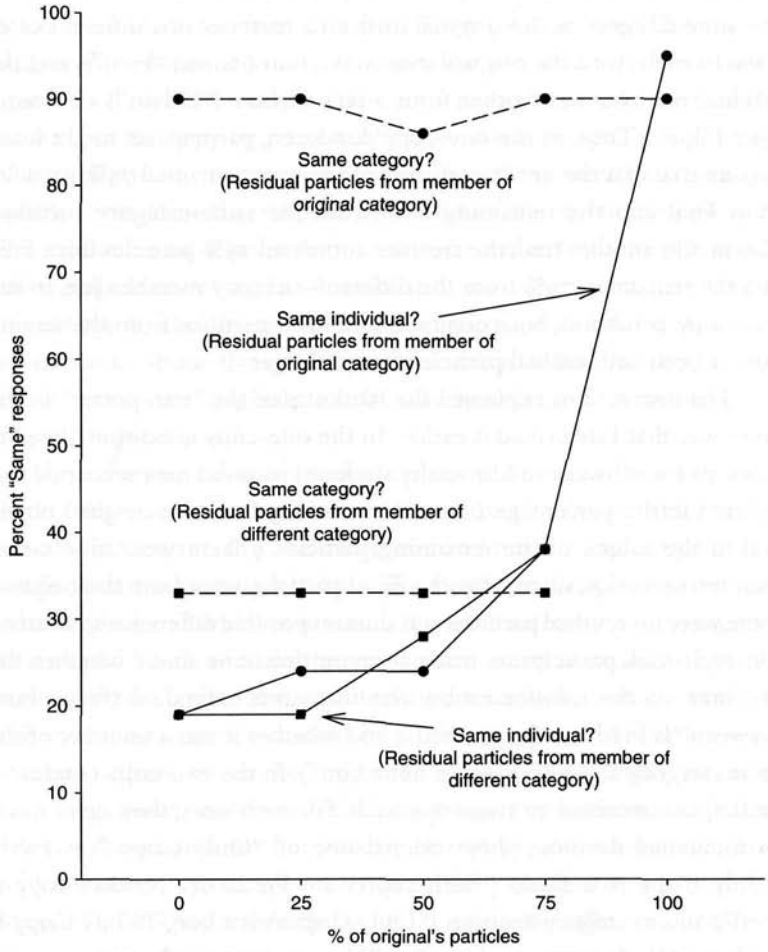

FIGURE 1.2 Percentage of responses indicating that the result of the transformation was still the same individual (solid lines) and was still a member of the same category (dashed lines). Lines with circles denote stories in which residual particles were from a member of the same category. Lines with squares indicate stories in which residual particles were from a member of a different category.

this figure indicates the percentage of the copy's particles coming from the original object, and the y-axis shows the percentage of trials on which participants agreed that the copy was the same individual as the original object (solid lines) or was in the same category as the original (dashed lines). Lines with circles represent stories in which the residual particles (those not copied from the original object) came from another member of the same species, while lines with squares are stories in which the residual particles come from a member of a different species. (For the two right-most points, all particles came from the original object, and there are no residual particles.)

Figure 1.2 shows that the larger the percentage of particles from the original individual, the more likely participants are to say that the copy is the same individual as the original. In the 0–75% range, the slope is fairly gradual but still amounts to an increase of 19 percentage points. There is no effect, though, of whether the residual particles are from a member of the same category or of a different category. By contrast, judgments of whether the copy is in the same category as the original produce the opposite effects. When the residual particles are from a member of the same category, participants agree that the copy is also a member of that category on 89% of trials. When any of the particles are from a member of a different category, however, agreement falls abruptly to 33% and does not vary with the proportion of particles from that category member.

The results in Figure 1.2 demonstrate that factors affecting category membership don't necessarily affect decisions about individual persistence. Although the source of the residual particles had a strong influence on category judgments, it had almost none on judgments of identity. This finding echoes the one I reported earlier (see Figure 1.1) and presents another puzzle for the view that persistence conditions come from knowledge of sortal membership. Assuming that "lion" is the relevant sortal, factors that cast doubt on whether the copy is a lion should also cast doubt on whether the copy is Fred, contrary to these results.

1.3.3 A Quantitative Version of the Causal Continuer Model

Results from the two-copy condition were similar to those from the one-copy condition in that judgments of individual identity depended

on the percentage of particles from the original individual, but not on the source of the remaining particles. Figure 1.3 graphs these results, with solid circles representing cases in which the residual particles were from a member of the same category (a different lion) and open circles representing cases in which the residual particles were from a member of a different category (a tiger). Each of the smaller graphs within Figure 1.3 corresponds to a combination in which one copy contained a given percentage of particles from the original individual (the initial lion) and the other copy contained another (possibly equal) percentage. For example, the graph in the lower left-hand corner represents the case in

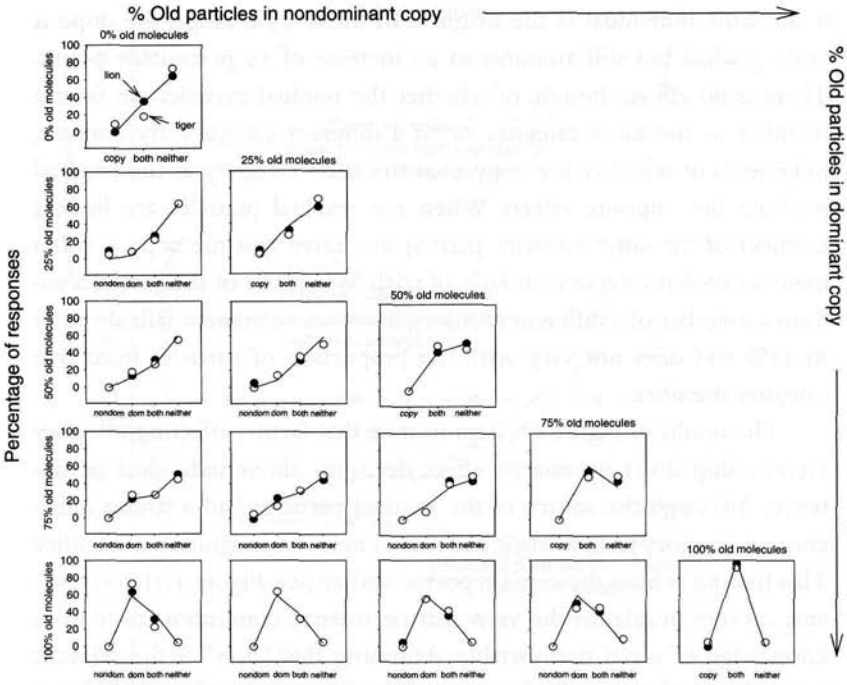

FIGURE 1.3 Percentage of responses that the dominant copy, nondominant copy, both copies, or neither copy was the same individual as the original. The graphs represent combinations in which each copy has either 0, 25, 50, 75, or 100% of its particles from the original object. Filled circles denote stories in which the residual particles were from a member of the same species. Open circles denote stories in which the residual particles were from a member of a different species. Lines are predictions from the Causal Continuer theory.

which one copy had 0% of its particles from the original individual and the other copy had 100% of its particles from the original. The points within each graph show the distribution of participants' responses. From left to right, these are the percentages of trials on which respondents judged: (a) that only the copy with fewer particles from the original (the *nondominant* copy) was identical to that original; (b) that only the copy with more particles from the original (the *dominant* copy) was identical; (c) that both copies were identical; and (d) that neither copy was identical. The graphs along the diagonal from the upper left to the lower right of the figure are cases in which both copies had the same percentage of original particles. In these cases, neither copy was dominant (and there were no other features to distinguish the copies); so I have combined the responses in which participants chose only one of these copies as identical. These responses are labeled *copy* on the *x*-axis. The solid lines in the graphs are the model's predictions, which I describe shortly.

Figure 1.3 highlights several trends in the results. First, the percentage of "dominant copy" or "copy" responses (relative to "both" or "neither" responses) increases from top to bottom, along the columns of graphs. The increase is steep between 75% and 100% of old particles, but is perceivable at lower levels as well. This indicates that as the percentage of original particles in the two copies becomes more dissimilar, participants shift toward thinking that only the dominant copy is identical to the original item. Second, a glance along the diagonal from upper left to lower right shows that the percentage of "both" responses increases (relative to "neither" responses). Both copies have the same proportion of original particles here, and as this proportion rises, participants increasingly believe that both copies are identical. Third, whether the residual particles came from a member of the same category as the original or from a different category has no effect on participants' choices. This finding replicates the results from the one-copy condition, as noted earlier. In applying the Causal Continuer model, I focus on these individual decisions. However, the decisions about category membership in the two-copy condition also replicate the one-copy condition in showing an effect of the residual particles' source, but not of the percentage of particles from the original. This echoes the dissociation in Figure 1.2.

The Causal Continuer approach is consistent with these trends. According to this theory, a participant's response on a particular trial

should depend on two decisions. First, she needs to determine whether one of the copies is causally closer than the other. Second, she also needs to know whether either copy is close enough to the original to qualify as identical to it. If the answer to both questions is "yes," she should respond that only the closer copy is identical. If the answer to the first question is "no" but the answer to the second is "yes," she should respond that both are identical. In all other cases (i.e., the answer to the second question is "no"), she should report that neither is identical. The first of these decisions is responsible for the increase in "dominant" responses along the columns of Figure 1.3. The greater the difference between the two copies in the percentage of original particles, the more likely the dominant copy is to be closer than the nondominant copy to the original, and the more likely participants are to make a "dominant" response. The second decision is responsible for the increase in "both" responses along the diagonal, where the two copies have the same percentage of original particles. The larger this percentage, the more likely "both" copies will be close enough, and the more likely participants are to make a "both" response. The lines in the Figure 1.3 graphs show the fit of a simple mathematical model based on combining these two decisions. The Appendix to this chapter contains the details of the model-fitting, but the results confirm the visual impression that the model does quite well, accounting for 96% of the variance with only a single free parameter.

1.3.4 An Experiment on the Effects of Causality and Similarity

It is reasonable to think that similarity between an object and its successor can sometimes provide evidence for identity over time. Similarity between Aunt Dahlia's appearance in 1970 and in 2011 may be enough to lead us to believe that these two manifestations belong to the same person. The Causal Continuer model claims, however, that causal factors can override similarity if the two factors conflict. We judge someone who is merely similar to Aunt Dahlia (but who is not a causal outgrowth of her earlier stages) as nonidentical, perhaps even as an imposter (for historical cases, see Barry, 2003; N. Z. Davis, 1983; Grann, 2008; Munsell, 1854). To see why, imagine an iceberg whose size is 3 x 3 x 3 m at a particular time t_0. Most people probably assume that over time icebergs tend to shrink due to temperature and to splitting (caused by stress from

storms and other factors).[12] Thus, at a later time t_1, the original iceberg's continuer would presumably have smaller dimensions rather than larger ones. The similarity of icebergs, however, might be more symmetric. The 3 x 3 x 3 m original might be about equally similar to a 4 x 4 x 4 m iceberg and to a 2 x 2 x 2 m iceberg at t_1, but only the latter is likely to be identical to the original.

To see whether causal beliefs do indeed dominate similarity, I asked participants in a further study to make judgments about icebergs.[13] In the experiment, participants read a scenario in which scientists were studying an iceberg named *Sample* 94, whose dimensions were 3 x 3 x 3 m. During the two parts of the experiment, I gave participants a list of icebergs of varying dimensions (e.g., 4 x 3 x 1 m or 2 x 1 x 1 m) that the instructions described as being found "sometime later" in the same vicinity. Participants rated both how similar each item was to the original Sample 94 and how likely the item *was to be* Sample 94. The goal of the study was to distinguish identity and similarity judgments. If causal mechanisms dominate judgments of identity, we should find that participants give lower identity ratings than similarity ratings to icebergs whose dimensions are greater than the original sample. Similarity and identity judgments may converge for icebergs whose dimensions are smaller than the original.

In their similarity and identity ratings, participants compared the 3 x 3 x 3 m iceberg to each of a set of items formed by combining the dimensions 4 m, 3 m, 2 m, and 1 m in all distinct ways. Thus, one item was 4 x 4 x 4 m, another 4 x 4 x 3 m, and so on. (The instructions told participants that the dimensions were always given with the larger sides first, without regard for the iceberg's orientation. For example, participants rated a 4 x 3 x 1 iceberg but not a 3 x 1 x 4 iceberg, since these would be the same item. Because of this aliasing, there were 20 items in the stimulus set, shown on the *x*-axis of Figure 1.4, below.) After each item was a rating scale, containing the numbers 0 to 9. I tested 46 Northwestern undergraduates in this experiment. Half these participants rated similarity first; half rated identity first.

When comparing the standard iceberg (3 x 3 x 3 m) to one with a larger dimension (e.g., 4 x 3 x 3 m), participants should see the second as potentially similar to the first but not identical to it. Because icebergs tend to shrink over time, a comparison iceberg with a larger dimension

can be similar but not identical to the standard. The mean ratings appear in Figure 1.4, and they confirm this prediction. Filled circles in the figure are mean identity ratings, and open circles mean similarity ratings. The *x*-axis lists the individual iceberg dimensions, with the vertical dashed line separating icebergs whose dimensions are all less than or equal to the standard from those icebergs containing one or more larger dimensions. When the comparison iceberg has a larger dimension (right side of the figure), its mean similarity rating is always higher than its identity rating, but when the comparison iceberg's dimensions are smaller or equal to those of the standard (left side of the figure), the ratings are more nearly equivalent.

As Figure 1.4 suggests, there is a significant interaction between type of judgment (similarity versus identity) and whether the iceberg has a dimension greater than that of the standard. We can get a more revealing

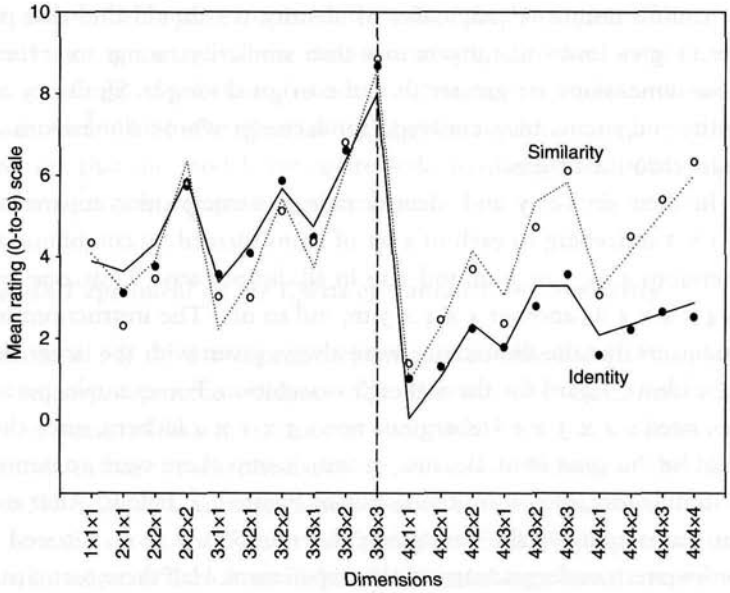

FIGURE 1.4 Mean ratings (0-to-9 scale) of similarity (open circles) and identity (filled circles) between icebergs of varying sizes (*x*-axis) and a 3 x 3 x 3 standard. The dashed line shows predictions for similarity ratings from a regression model; the solid line shows predictions for identity ratings (see text for a description of these models).

picture, though, by examining variables that may have contributed to participants' reasoning about these judgments. Figure 1.4 shows peaks in the ratings when the icebergs were cubical (e.g., 2 x 2 x 2 or 4 x 4 x 4) or nearly so (e.g., 3 x 2 x 2 or 4 x 3 x 3), suggesting that participants were taking into account the iceberg's shape. Because the standard iceberg was itself cubical, participants may have given the comparison iceberg higher ratings if it too had approximately the same shape. In addition, participants considered overall size, giving higher ratings when the size of the comparison iceberg was nearly that of the standard. Initial analyses hinted that participants might have compared the icebergs in terms of the sum of their dimensions rather than their product, possibly for computational ease. Icebergs whose dimensions summed to a total near 9, the sum of the dimensions of the 3 x 3 x 3 standard, got higher identity and similarity ratings than the others. Compare, for example, the ratings of the 4 x 3 x 2 item (dimension sum = 9) to the 4 x 2 x 1 item (dimension sum = 7) in Figure 1.4.

To see how well these factors predicted the mean ratings, I applied two regression equations to the Figure 1.4 data, one for the similarity and the other for the identity judgments. Both equations contained three terms: The first captured departure from cubical shape in terms of the standard deviation of the iceberg's three dimensions. The second term measured the overall difference in size between the standard and comparison iceberg, using the sum of the dimensions, as just discussed. That is, if d_1, d_2, and d_3 are the dimensions of the comparison iceberg, then the value of this term was $| d_1 + d_2 + d_3 - 9 |$. The final term was a binary indicator of whether any of the iceberg's dimensions was greater than that of the standard (1 if one or more of the dimensions was 4 m, and 0 otherwise). I expected this last term to discriminate the identity ratings from the similarity ratings. Predictions from these two regression equations appear in Figure 1.4: The solid line corresponds to the identity predictions and accounts for 95.4% of the variance among the means. All three terms produced statistically significant coefficients. Figure 1.4 shows the predictions for the similarity ratings as the dotted line, and these predictions account for 92.6% of the data. The shape and size terms were significant in this analysis, but as we would expect, there was no effect on the similarity ratings of whether the comparison iceberg contained a dimension larger than those of the standard.

When one of the comparison iceberg's dimensions exceeded the standard's, participants discounted the possibility that it could *be* the standard but not the possibility that it could be similar to it. Thus, people's judgments of an object's identity are not simply a matter of similarity. They involve the causal trajectory of the item as it evolves—in this case, shrinking rather than stretching over time.[14]

1.4 Fission and Fusion

Although I believe that the Causal Continuer model has advantages over earlier approaches, we've considered so far only a fairly narrow range of identity judgments. I've focused on situations that are difficult for other theories to explain—ones that deliberately eliminate spatiotemporal continuity (destroying the particles of the original object in the lion scenarios) and that dissociate identity from basic-level category membership and similarity. We should also ask, however, whether the model can deal with other identity issues. In this section, I consider two further examples of identity questions from previous research.

1.4.1 The Ship of Theseus

One famous test case for theories of identity is due to Thomas Hobbes (1838–1845) and is the subject of some recent research in developmental psychology (Hall, 1998; Noles & Bloom, 2006).

> *The Ship of Theseus:* A wooden ship was repaired over a long interval by removing individual planks one-at-a-time and replacing them at each step with new ones. This process continued until none of the old planks remained, and the ship consisted entirely of new planks. However, the old planks were stored and then reassembled exactly as before. Two ships exist at this later point, each of which could claim to be the original ship: the one with old planks and the one with new planks. Which, if either, is Theseus's ship?

The Causal Continuer model can afford to be neutral with respect to this question (see Nozick, 1981). Both resulting ships—call them *Old Parts*

and *New Parts*—are causal outgrowths of the original. This is a case of fission, in which an initial object gives rise to two possible successors. In this case, New Parts enjoys closer temporal continuity with the original, while Old Parts has greater overlap in material composition. Whether we deem Old Parts, New Parts, both, or neither as Theseus's ship will then depend on how we weigh these two factors. The model does not make an a priori decision among the options, but it does explain the uncertainty we feel about the choice. Both composition and temporal overlap are typically important and perfectly confounded in identity judgments about ordinary ships. Both are diagnostic of the causal forces that support a ship's existence. Hobbes's story separates these factors, forcing us to consider them independently, and this demand for independent weighting creates the puzzle. In the same manner, the model accounts for the intuition that either Old Parts *or* New Parts would unambiguously be Theseus's ship if the other were out of the picture. For example, if the original ship were simply disassembled and reassembled, we probably wouldn't hesitate to identify it with the ship of Theseus. Similarly, if the parts of the original were gradually replaced with no reassembly of the old parts, the ship of Theseus would be the repaired ship. What creates indecision in Hobbes's problem is the competition between Old Parts and New Parts for being the closest or best option. The same factor produced the equivalent effect in the multiple-copy experiment I described earlier.

A study by Hall (1998) provides some evidence about children's and adults' preferences in a closely related problem. In Hall's version, participants heard stories and saw pictures of a star-shaped object, called "Sam's quiggle," which the stories described as an artifact (a kind of paperweight) in one condition and as a natural object (a jungle animal) in another. This object loses its old parts and gains new ones until it is composed entirely of new parts. In some stories, a person performs this substitution; in others, no agent is specified (the change "simply happens"). Someone then reassembles the old parts as before. The participants had to choose whether Old Parts or New Parts was Sam's quiggle. The results showed that adults tended to choose New Parts when the object was an animal that lost and gained parts spontaneously, and they chose Old Parts when the original object was an artifact that a human revamped. In the two remaining cases (animal that a human revamps and artifact that

changes spontaneously), adults split their vote. Five year-olds favored Old Parts in all conditions, with seven year-olds showing a pattern intermediate between that of younger children and adults.

We can't directly apply the Causal Continuer theory to Hall's results because we don't have an independent measure of the causal closeness of the original to the two resulting objects. The theory is consistent with the pattern of data, however, given two reasonable assumptions. The first of these concerns adults' biological knowledge about the transformations. Adults know that living things, unlike artifacts, rarely survive complete disassembly. In operating on a live animal, for example, a surgeon must be careful to keep most of the animal intact if it is to survive the operation. By contrast, persistence over complete disassembly and reassembly is much more plausible in the case of an artifact. Disassembling a multipart paperweight and putting it back together can produce a perfectly good paperweight. This distinction between natural objects and artifacts would tend to shift adult responses toward New Parts in the natural-object condition, since Old Parts would no longer be a living creature. Younger children may lack such information and may therefore treat natural objects like artifacts in this respect.

Biological knowledge, however, is not sufficient to explain all facets of the data. For example, when the object was an artifact and a person replaced each part to create New Parts, both adults and children overwhelmingly favored Old Parts. This condition is the one most similar to the standard Ship-of-Theseus puzzle, where opinion seems more evenly divided between the two contenders. Why then did participants in Hall's study regard Old Parts as the identical item? One possibility concerns the details of the change. When an agent performed the substitution creating New Parts, the stories described the change as occurring over a several week period, with the agent replacing one part per week. This discontinuous change may have weakened the causal link between the original object and New Parts, causing participants to choose Old Parts instead in both the artifact and natural object conditions. The pictures illustrating the individual steps may have abetted the feeling of discontinuity by showing stages in which parts were missing from the object that was eventually to become New Parts. By contrast, when there was no human agent responsible for the change, participants probably saw the

transformation that produced New Parts as more continuous (e.g., a kind of molting), as Hall (1998) suggests.

Taken together, these assumptions explain the results in terms that are congenial to the Causal Continuer model. Although the assumptions are obviously after-the-fact, they seem plausible and provide a bridge between Ship-of-Theseus cases and the new results described here.

1.4.2 Fission and Fusion in Memory

Our concept of an individual object can sometimes undergo fission or fusion, even when the object itself is unchanged. I had read from time to time about a remarkable British polymath, Sir William Hamilton, who, among other accomplishments, was a mathematician (W. R. Hamilton, 1866/1969), astronomer, expert on volcanoes (the subject of Susan Sontag's, 1992, novel, *The Volcano Lover*), diplomat, collector of antiquities, and philosopher (the target of Mill's, 1868, *Examination of Sir William Hamilton's Philosophy*). I took a surprisingly long time to realize that this individual was really three different people—an astronomer-mathematician, a volcanologist-diplomat-collector, and a philosopher—each named "Sir William Hamilton." This discovery meant creating new singular concepts and reassigning the properties of the old merged Hamilton concept to its fissioned counterparts.

The opposite process, conceptual fusion, sometimes also occurs in revising our knowledge of people. I might have learned about Art Jones, the softball coach, in one context, and Arthur P. Jones, the Chevy salesman, in another, and only later determined that these two Joneses were the same. Fusion cases like this are analogous to the problem originally described by Frege (1892/1952) for the meaning of identity statements—for example, the Morning Star is identical to the Evening Star. To put this issue in a more contemporary light, suppose the meaning of a singular concept like ARTHUR JONES is a particular individual, Jones himself. Many philosophers currently believe that this meaning is fixed by a causal-historical connection that runs from the denoted individual (Jones) to the person who possesses the concept (me). But since only one Arthur Jones exists, who is both the softball coach and the Chevy salesman, my ARTHUR JONES concepts must have referred to him all along.

So how can it be a surprise for me to discover that only one individual is involved rather than two? (See Jeshion, 2010; Lawlor, 2001; and Perry, 2001, for recent philosophical treatments of this issue.)

John Anderson (1977; J. R. Anderson & Hastie, 1974) has studied the fusion case by presenting participants with identity information before or after they had learned separate facts about individuals. In one condition, for example, participants first learned *James Bartlett played the banjo* and *The lawyer sold the boat* (among other unrelated sentences) and then learned *James Bartlett is the lawyer.* Response times to verify directly stated information (*Bartlett played the banjo*) versus inferred information (*Bartlett sold the boat*) suggested that participants transferred the predicates originally associated with the proper name to the concept associated with the description. For example, the predicate *played the banjo* would come to be directly associated with the concept THE LAWYER and the concept BARTLETT would be abandoned (J. R. Anderson, 1977).

Anderson (1977) suggests that which concept is retained and which abandoned might depend on the relative amount of information connected to the two. We're more likely to retain a concept that is associated with more information, since less work is then required in revising memory. According to the present perspective, however, the revision process might also depend on relations between the old concepts and the revised ones. In fusion cases like Anderson's, we might prefer to keep the concept of the person that we can most easily imagine becoming the merged individual, the person who could more readily acquire the properties of the other. This may be the individual who already has more properties, in line with Anderson's hypothesis, but may also be the one whose properties are less malleable, more reliable, or fixed over a longer interval. In these cases, a causal transition to a merged state is easier to envision. Although I know of no direct test of this hypothesis, it seems consistent with other examples of imagined change (see, e.g., Kahneman & Miller, 1986, and the discussion of counterfactuals in Chapter 3 of this book).

In the case of conceptual fission, like the initial Hamilton example, the Causal Continuer approach likewise suggests that the conceptual change may be similar to what would happen if actual fission were to take place. I had to create new representations for some of the new people that I discovered in order to have separate concepts for each of

the Hamiltons. One possibility is to abandon completely the old Hamilton concept and create three new ones for each of the "descendents." This is analogous to a "neither" response in the lion experiment, and it might occur if none of the true Hamiltons is more closely related than the others to the old false concept. But an alternative is to retain a version of the old merged concept, editing it to represent one of the final individuals, and then construct just two new representations for the others. This is analogous to choosing one of the potential continuers as identical to the original in actual fission cases. If we can easily imagine a causal transition from the merged individual to one of the final people, then we might reasonably choose to modify the old concept to represent him, constructing new representations for the others. For example, if we can conceive the old merged Hamilton shedding some of his properties to become the diplomat-volcanologist-collector, then it might be easiest to modify the original concept to represent that person and create new concepts for the mathematician-astronomer and for the philosopher.

Conceptual versions of causal-continuer effects may also influence the ease with which people can track characters in stories or assign referents to anaphoric expressions, such as pronouns or definite noun phrases. (See, e.g., Caramazza, Grober, Garvey, & Yates, 1977, for evidence of effects of causal prominence on pronoun assignment, and Rudolph & Försterling, 1997, for a review.)

1.5 Extensions and Limitations

The Causal Continuer approach contends that a later manifestation of a single object must causally stem from earlier ones, so that causality takes precedence over qualitative overlap in properties, spatiotemporal continuity, or sortal membership. Similarity, continuity, and other properties can come into play, however, if direct causal information is absent or ambiguous. The model makes its identity judgments on the basis of two interrelated decisions: An object x_0 is identical to another x_1 if x_1 is causally close enough to be the continuation of x_0 and if x_1 is the closest of all the close-enough competitors. This is the answer the model gives to our opening question in (1). Evidence for this approach comes from the study in Section 1.3.2, which manipulated the closeness of

an original object to each of two possible continuers. The model succeeded in predicting participants' decisions about which continuer was identical to the original in a setting where there was no spatial continuity between the items. The same study showed a dissociation between these identity judgments and judgments of basic-level category membership. The study in Section 1.3.4 tested the model's prediction that people rely on causal continuity over similarity when these two factors are at odds.

The model appears to describe participants' responses in these experiments, but some of the model's principles may seem puzzling for theoretical reasons. First, the model goes along with judgments that a single object can divide into two and remain identical to both descendants. Such judgments seem to imply that object identity is intransitive, contrary to property (2c), and they raise issues about whether the model's (and the participants') concept of identity is coherent. Second, the model maintains that object identity is not necessarily tied to the category-level concepts to which the object belongs. However, there are some presumptive reasons to think that categories must be involved in any identity decision. The rest of this chapter considers ways to resolve these two problems.

1.5.1 Transitivity of Identity

Although the Causal Continuer model provides a good account of the data from the dual copy experiment, this accomplishment depends on its liberal policy with respect to the "both" responses. The model produces these responses when the difference between the possible continuers is small enough to be ignored. In this case, if either copy is close enough to be identical, then both must be identical; otherwise, neither is identical. This assumption is consistent with participants' responses: In the case in which both contenders consisted only of particles copied from the original, nearly all participants made a "both" response (see the bottom right graph in Figure 1.3). The trouble is that these responses appear to violate the transitivity property of identity in (2c). How could both copies be identical to the original while not being identical to each other?

We could view these responses as mistakes in participants' thinking about identity. Perhaps participants' identity decisions reflect a simple

heuristic rather than a considered, normatively appropriate procedure. For example, they may have used the causal distance between the original item and the copies, without concerning themselves with the extra constraints that identity imposes. This behavior might be the result of the relatively greater importance of causal continuity over strict identity in dealing with issues of survival and persistence (as Parfit, 1984, argues; see also Bartels & Rips, 2010, for evidence supporting Parfit's position). According to this approach, the responses are much like intransitivities in the preference judgments of decision makers (Tversky, 1969): Sometimes people prefer option A over option B and B over C, yet prefer C over A.

Alternatively, we could interpret the experimental findings in a way that brings the "both" responses in line with transitivity. We have been assuming that participants believe the two copies in the experiment are distinct individuals, and this assumption leads to intransitivity when both copies are also identical to the original. But another way of viewing the situation is that the transporter produces, not two independent objects, but two parts of a single temporally branching one (this is one of the individuals or "lifetimes" that Perry, 1972, describes in fission cases). Figure 1.5a schematically illustrates this approach. The diagram indicates

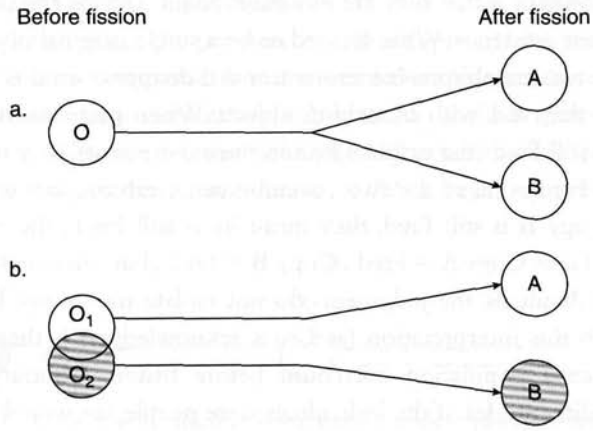

FIGURE 1.5 Two ways of interpreting fission examples. (a) The original and the copies are temporal and spatial parts of a single branching object, and (b) each copy is a distinct object that overlaps with the other spatially during the initial stage of its life and diverges thereafter.

the temporal sequence of events in the life of the lion, from its birth at the left-hand side, to the point at which it is copied, to its end state at the right. According to this way of thinking, the duplicated lion in the stimulus stories exists after division in something like the way that a tree exists spatially in its multiple branches. Just as the branches are parts of the same tree, the multiple copies are parts of the same creature. No intransitivity appears on this interpretation: Copy A, copy B, and the original object, O, are all the same individual.

A possible problem for this solution, however, is that it is difficult to shake the idea that the two copies must be nonidentical, since each can presumably function on its own, develop distinct properties, and appear and behave just like two ordinary objects, despite their common origin. According to this counterargument, the copies are more like identical twins (or embryonic clones) than like a single temporally branching object. Although they have a common origin, identical twins count as two people in a census, have two votes in an election, and so on.

A second way of salvaging transitivity is to construe the two copies as distinct objects, but ones that existed all along, sharing the spatial parts of the original (Lewis, 1983). Figure 1.5b illustrates this reinterpretation. Copy A begins life when the original does, surviving the division and continuing on its own way. Copy B does the same. What's unusual about these individuals is that they are indistinguishable during the pre-fission part of their existence: What seemed to be a single original object turns out to be two cohabitors. Intransitivities also disappear on this interpretation, as they did with branching objects: When participants say that copy A is still Fred (the original lion in the experiment), they mean that he is still $Fred_1$, one of the two co-embodied creatures, and when they say that copy B is still Fred, they mean he is still $Fred_2$, the other co-embodied one. Copy A = $Fred_1$, Copy B = $Fred_2$, but since these two are distinct individuals, the judgments do not violate transitivity. The difficulty with this interpretation (as Lewis acknowledges) is that it seems to produce a population overcount before fission. Contrary to the co-embodiment idea, if the individuals were people, we would probably refuse to count the pre-fission stage twice in a census, would deny it two votes in an election, and so on. Moreover, although this alternative keeps participants' judgments from being inconsistent with the identity axioms, it does so at the cost of positing a hidden ambiguity in the proper name

for the original object in the two-copy condition (*Fred* can refer to either $Fred_1$ or $Fred_2$), where no such ambiguity was present in the one-copy condition.

Both these solutions to the transitivity problem come at a high price. Positing co-embodied individuals appears to produce too many pre-fission objects, while positing branching individuals produces too few post-fission ones. Although there are ways of reconciling these solutions with our instinctive ways of counting, they require adjustments to our counting strategies (e.g., counting object stages rather than objects, as in Lewis, 1983). Still, these distinct ways of interpreting fission cases stand as alternatives to the view that participants were committing a performance error or making a mistake in judgment. "Both" responses rule out some ways of construing the two-copy condition; they eliminate the possibility that the initial object has gone out of existence and two new ones have appeared. Nonetheless, they leave open other possibilities that could be explored, such as branching or co-embodiment. Which interpretation is correct is an issue that must remain open here.

In some respects, the participants' situation parallels that of observers in certain types of apparent motion experiments (Ullman, 1979). Figure 1.6 illustrates the simplest situation of this type. An observer sees a central dot, x_0, in an initial display. This dot disappears, and then two dots, x_1 and x_2, appear in a second display, with x_1 and x_2 located on either side of, and equally distant from, the position x_0 had occupied. If the interstimulus interval is appropriate and the observer fixates x_0, then he or she sees simultaneous movement toward both x_1 and x_2, as the arrows indicate in Figure 1.6. However, on the assumption that motion correspondence implies identity (Kahneman, Treisman, & Gibbs, 1992), we get a potential

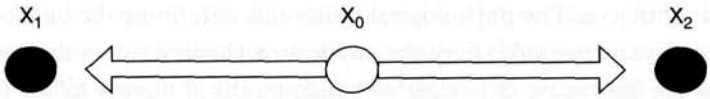

FIGURE 1.6 A situation in which apparent motion leads to perceived fission of an object. A display containing dot x_0 is presented first. This dot disappears and is followed by a second display containing x_1 and x_2 at the locations shown. In this case, x_0 appears to move simultaneously to both x_1 and x_2. (Arrows show the direction of motion but do not appear in the display.)

violation of the transitivity relation in (2c). If $x_1 = x_0$ and $x_0 = x_2$, then transitivity yields $x_1 = x_2$. But it appears that $x_1 \neq x_2$ since these two dots are in separate locations in the second display. However, are observers who report motion in opposite directions committing a performance error? Committing themselves to the idea that identity is intransitive? We can only interpret the report as an intransitivity error if we reject alternative interpretations of these judgments, such as branching.

1.5.2 Relations to Earlier Theories

The Causal Continuer approach seems capable of handling many of the issues that created obstacles for earlier theories. Because the model subordinates similarity judgments to causal continuity, it explains why similarity can function as evidence for identity in some situations but as evidence against it in others. For example, a difference in size (a dissimilarity) may support the hypothesis that the flower you perceive now is the same one you planted earlier, but contradict the hypothesis that the cup you perceive now is the same one you washed earlier.

Along the same lines, although knowledge of spatiotemporal continuity is an important clue to sameness, it need not be decisive. In the vicinity of a dual-ing machine, for example, causal facts about the device blocks the inference from continuity to the conclusion that the later object is a causal outgrowth of the earlier. We needn't resort to any kind of spatiotemporal continuity if we already know the causal facts.

The Causal Continuer theory assumes that (people believe that) causal forces (and the objects they create) exist in their own right, independent of language and thought. In particular, physical objects don't depend on the concepts or categories to which these objects belong. Of course, different kinds of causes may support the existence of different kinds of objects. The physiological causes that determine the life course of cats or canaries differ from the physical-mechanical causes that determine the life course of bridges or buildings. But it doesn't follow from this difference in type of cause that objects inherit their identity conditions from their sortal categories.

All theories of identity must acknowledge that objects vary in their behavior in ways that are important for identity and persistence. Dropping a wine glass on a slate floor from a height of 3 ft. will probably cause it

to shatter and go out of existence, whereas dropping a cat on the same floor from the same height will probably leave it unscathed. But this domain specificity does not distinguish between the sortal and the Causal Continuer approaches. What does distinguish the theories is the explanation for such differences. In the case of the sortal view, the source of the differences is the meaning of the sortal terms that describe the objects. Part of the meaning of *(wine) glass*, for example, is an identity condition (see (4) above) that stipulates that nothing following a shattering event can be identical to the original glass. By contrast, the Causal Continuer theory accounts for the difference in terms of the kinds of causes responsible for maintaining the integrity of the object in question. It is an empirical fact, and not part of the meaning of *glass* or *cat*, that some of the causes that disrupt a glass's existence do no damage to a cat.[15]

An analogy may make this distinction clearer. The internal temperature of objects varies by domain, with some types of objects having systematically higher temperature than do others. The body temperature of birds, for example, tends to be higher than that of humans under normal conditions. In a sense, then, body temperature could be said to be "sortal relative." But no one would suppose that the meanings of the terms *bird* and *human* include "temperature conditions" that specify the allowable range of body temperatures in these species. Instead, the temperature of different creatures is the result of mechanisms of thermal regulation, among other causal factors. In a parallel way, the Causal Continuer theory claims that domain differences in identity are due to differences in the kinds of causal mechanisms that maintain an object during its career rather than to differences in the meaning of expressions for these objects.

To see that sortals are not necessary, notice that examples of sortal-relative identity conditions are in short supply. Sortal theories need these conditions to specify the R_S relations in (4). But no clear examples of identity conditions exist for everyday sortals such as *cats* or *trucks*, with the possible exception of (much disputed) criteria for *persons*. What are the necessary and sufficient conditions that cat x at t_1 and cat y at t_2 must possess in order for x to be identical to y [i.e., what is $R_{CAT}(x,y)$], and how do they differ from those conditions for dogs [$R_{DOG}(x,y)$]? (See Mackie, 2006, for similar complaints.) The difficulties for sortal theories of singular concepts parallel the well-known difficulties for classical

theories of category-level concepts (see Murphy, 2002; E. E. Smith & Medin, 1981). There are few convincing examples of necessary and sufficient properties for membership in everyday categories, and most cognitive psychologists have given up hope of uncovering them. We suspect that in the case of sortal theories, too, the shortage of plausible examples is due to the fact that people simply do not know sortal-relative conditions of identity for everyday categories.[16] If so, and if sortals are count nouns that furnish identity conditions, then there are few or no sortals.

1.5.3 Identity and Modal Thinking

I mentioned at the beginning of this chapter that our concepts of people and other things must be rich enough to support conjectures about what might have happened to these individuals in situations that are possible but never actually take place. The concept of my friend Georgine, for example, informs my guesses about how she will behave in settings she hasn't yet, and perhaps never will, experience. The same goes for predictions about political figures or celebrities whose dispositions I think I know. In the realm of inanimate objects, predictions about location and change have the warrant of well-established physical principles, even when the predictions' initial conditions never occur. What give us the ability to make these counterfactual judgments are the same causal relations that, according to the Causal Continuer theory, govern our ability to trace these individuals in the real world. As many theorists have argued, causal relations yield law-like generalizations that support our theorizing.

To see the similarity between counterfactual judgments and judgments of identity, consider the relation between an ordinary historical narrative and a historical fiction. Both stories might begin with the same set of events—for example, the actual events that have occurred during the life of Georgine from her birth in 1950 to her 30th birthday in 1980. The straight historical account would continue to follow the actual causal stream from 1980 to the present, but the historical fiction might diverge from the true state of affairs, perhaps beginning with a fictitious chance event that Georgine is said to experience in 1980. The author of the historical fiction could then elaborate the counterfactual post-1980 story by spinning out the causal consequences that follow the fictitious event and the actual Georgine events that preceded it. This elaboration

might require adjusting these actual events in order to accommodate the fictitious ones. "Minor miracles" may be necessary to explain the divergence (Lewis, 1979), but a plausible story would make such adjustments in a way that minimizes changes to the facts. However, both the factual and fictional narratives make use of most of the same causal principles that get Georgine from one moment on her time line to the next. When we try to imagine what Georgine would be like if such and such a counterfactual event had taken place, these principles organize our projections.

Category-level concepts display many of these same normative features. We can reason about what will happen to categories under unknown or counterfactual conditions, drawing out the consequences, for example, of supposing that poodles can bite through wire (e.g., Osherson, Smith, Shafir, Gualtierotti, & Biolsi, 1995) or that furniture is eaten at the end of a meal (Sternberg & Gastel, 1989). We can also state generalizations about these categories (e.g., that lions have manes) that withstand numerous exceptions (female or immature lions). It seems likely that causal knowledge about these categories is again responsible for these abilities. These facts raise issues about people's ability to represent causal relations and about the difference between relations that sustain individuals and those that sustain categories. Chapters 3 and 4 in this book are devoted to these questions, but before tackling them, I'd like to consider some further questions about identity and individuation that arise in mathematical contexts.

Appendix

A Mathematical Version of the Causal Continuer Theory

To fit the Causal Continuer model to the data in Figure 1.3, we can assume that causal closeness in this experiment depends on the percentage of the copy's particles that derives from the original. In the stories, the "transporter" is the causal mechanism that produces closeness by copying particles and transmitting them. We might therefore represent the probability that the dominant copy, *d*, is closer than the nondominant copy, *n*, in terms of the ratio in (A1), when the proportion of original particles in *n* is less than 1:

$$(A1)\ \Pr(d\ closer) = \frac{k \cdot (proportion\ original\ particles\ in\ d - proportion\ original\ particles\ in\ n)}{1 - proportion\ original\ particles\ in\ n}$$

When the proportion of original particles in *n* is 1, we can define Pr(*d* closer) = 0. In Equation (A1), *k* is a free parameter representing the maximum probability that *d* can attain. Even if copy *d* has all its particles from the original and *n* has none, some participants might still feel that there is not enough difference between them for *d* to be causally closer than *n*. We can also assume that if *d* is not closer (with probability 1 − Pr(*d* closer)), then we have a tie (i.e., *n* can never be closer than *d*).

To predict the data, we must also determine whether either copy is close enough to be potentially identical to the original item. Since the same participants made identity judgments for each copy separately in the one-copy condition, we can use these decisions to estimate empirically the likelihood of a "yes" answer to this question. During one of the two-copy trials, for example, participants learned that one copy contains 75% of its particles from the original and the second copy contains 25%. Participants had judged that a copy with 75% original particles was identical to the original on .38 of trials and that a copy with 25% original particles was identical on .21 of the trials in the one-copy condition. We can then estimate the likelihood that one or the other is causally close enough to be identical as $1 - (1 - .38)*(1 - .21) = .51$. The general relationship is that in (A2):

$$(A2)\ \Pr(d \text{ or } n \text{ close enough}) = 1 - (1 - \Pr(d \text{ close enough})) \cdot (1 - \Pr(n \text{ close enough})).$$

Combining Equations (A1) and (A2) gives us the predictions for the two-copy condition in Figure 1.3. For example, Pr(*d* closer)·Pr(*d* or *n* close enough) is the probability that participants should identify only the dominant copy as identical to the original. Similarly, (1 - Pr(*d* closer))·Pr(*d* or *n* close enough) is the probability of a "both" response. To evaluate the model, I fit these equations to the data in Figure 1.3, using nonlinear least-squares approximation. Since there is no apparent difference between cases in which the residual particles were from the same or different species, I collapsed the data from these two conditions before fitting the model. As noted earlier, the model predicts that participants should never respond that only the nondominant copy is identical to the original. Figure 1.3 shows that this is approximately true, but I omitted these points in fitting in order to obtain a more conservative view of the model's accuracy. The model was therefore fit to 45 data points: the "dominant only," "both," and "neither" responses in the 15 graphs in Figure 1.3. The resulting predictions appear as the lines in the figure, and the overall fit of the model is quite good. The root mean square deviation (*RMSD*) for the 45 critical observations is only 5.1 percentage points, and $R^2 = .957$. The value of the single free parameter, *k*, from Equation (A1) is 0.62.

Another way to evaluate the model is to compare it to a simpler variant. Suppose, for example, that participants make their decisions based on their separate judgments of whether the dominant copy is identical and whether the nondominant copy is identical. This procedure differs from the Causal Continuer idea in that there is no explicit comparison for closeness of the sort embodied in Equation (A1). If we represent the probability that the dominant copy is close enough to be identical as Pr(*d* close enough) and the probability that the nondominant copy is close enough as Pr(*n* close enough), as we did in (A2), then the probability that both are identical should be Pr(*d* close enough)·Pr(*n* close enough), assuming independence between the decision. Similarly, the probability that only the dominant copy is identical is Pr(*d* close enough)·(1 - Pr(*n* close enough)), and so on. Estimating the component probabilities from the one-choice data, as we did earlier, allows us to fit this simpler model directly with no free parameters. This model does considerably less well than the one I have just described (*RMSD* = 16.1 percentage points and R^2 = .618). The discrepancy is especially marked for "both" responses when the proportion of original particles is the same in the two copies, since the simpler model greatly underpredicts these proportions. In this model, a "both" response depends on both copies being independently close enough to be identical, as just noted. In the full model, however, there is no relevant difference between the two copies (the value of Pr(*d* closer) = 0 in Equation (A1)); so a "both" response depends on whether *either* copy could be considered close enough, as given by Equation (A2). This is typically a much larger value, in accord with the data. A likelihood ratio test (Bates & Watts, 1988) shows that the Causal Continuer model significantly improves on the simpler model, taking into account the former model's extra parameter.

NOTES

This chapter is based on an earlier article with Sergey Blok and George Newman (Rips, Blok, & Newman, 2006). I've also taken some material from Blok, Newman, and Rips (2007). In addition to Serge and George, I thank Jennifer Asmuth, Dan Bartels, Jennifer Behr, Amber Bloomfield, Aveen Farooq, Robert Goldstone, Gabe Greenberg, Douglas Medin, Ariela Lazar, Beth Lynch, Jeff Pasch, Andrea Proctor, Eyal Sagi, Steven Sloman, Elizabeth Spelke, Edward Smith,

and Sandra Waxman for their help on the earlier versions of this chapter. Some of the ideas developed in classes on object identity at Northwestern University, and I thank the students in these classes for their suggestions.

1. The relation between object identity and traditional recognition memory may not be straightforward. The standard recognition task is in some ways more about categorization than about object identity. If you were presented with the word *eggplant* and are now asked whether it was on an earlier list, the correct answer is "yes" even if the word now appears in a different font, color, or modality. The correct answer depends on whether the original word and the current word are tokens of the same type, but as I have already indicated, identity judgments are decisions about whether two appearances belong to the same token (i.e., are numerically identical). The relationship between perceptual object recognition and judgments of identity is potentially much closer. But even here, much of the research on object recognition is devoted to how people recognize objects as members of categories (e.g., horses) rather than on how they identify individuals (see Peterson, 2001, for a review of theories of object recognition). For example, the announced goal of Biederman's (1987) recognition-by-components theory is "to account for the initial categorization of isolated objects. Often, but not always, this categorization will be at a basic level, for example, when we know that a given object is a typewriter, a banana, or a giraffe." This is not to say that recognition is irrelevant to judgments of object identity, but only that the relationships need to be carefully worked out.

2. It is possible to debate whether the computer exists during the time at which it is disassembled. Whether people view a disassembled object as the same individual may depend on the extent of the transformation (e.g., the number of resulting pieces or the size of these pieces). For instance, people may be more likely to believe that a scattered collection consisting of the disassembled top and legs of a table is still the same individual than a scattered collection consisting of the zillions of disassembled circuit components of a computer (see Gutheil, Bloom, Valderrama, & Freedman, 2004, for relevant evidence). If the computer does not exist when its components are disassembled (as seems likely), then the example shows that objects can survive gaps in time. But even if the computer continues to exist during its disassembled phase, it clearly doesn't exist as a spatially continuous entity. Therefore, transformations can preserve identity across (at least) spatial discontinuity.

3. The term *sortal* is due to Locke (1690/1975, p. 417) in the same famous passage in which he distinguishes real and nominal essences. Wiggins (2001) points out that Locke's, Strawson's, and his own use of *sortal* derive from Aristotle's distinction between categories of substance and qualities.

4. Sortal theories in psychology appeal almost exclusively to principle (5), as we will see later in this chapter, but it is very difficult to state this

principle adequately. The main problem is that some sortal theories allow sortal categories to be nested. According to Xu (1997), for example, both *dog* and *physical object* are sortals with distinct identity conditions, R_{DOG} and R_{OBJECT}. Hence, Fido can go from being a dog to being a non-dog as long as he is covered by the sortal *physical object*. If we can always appeal to *physical object* as a sortal, however, then ordinary objects cannot go out of existence without somehow becoming nonphysical. This is inconsistent with the intuition that a chair that is splintered by an axe ceases to exist rather than continues to exist as a pile of wood scraps. I'm unsure whether there is a way to formulate (5) that is not question-begging, but we can safely leave this problem for proponents of sortal theories.

5. As mentioned in the Introduction to this book, I follow the usual convention of spelling names for concepts in all caps and names for linguistic entities (e.g., words or sentences) in italics or quotation marks.

6. One possible issue, and a source of conflict with sortal theories in philosophy (e.g., Wiggins, 2001), is that sortals like *cup* or *elephant* should also be necessary in order to individuate objects that appear together in the perceptual field. The evidence from Xu and Carey's experiments (Xu & Carey, 1995; Xu et al., 2004), however, is that younger infants do perform correctly when they have the advantage of previewing the objects. To explain this difference in performance, Xu and Carey argue that even the younger infants have a high-level sortal concept, equivalent to the concept PHYSICAL OBJECT, that Spelke has posited to explain infants' object tracking (e.g., Spelke, 1990; Spelke, Gutheil, & Van de Walle, 1995). This concept provides the sortal information that infants use in the preview condition. As Xu (1997, p. 369) states, "for both adults and young infants, there is nonetheless a sortal *physical object*, which is more general than *person*, *car*, or *tree*. A physical object is defined as any three-dimensional, bounded entity that moves on a spatiotemporally continuous path" (see also, Carey, 1995a; Carey & Xu, 1999). But sortal theories in philosophy typically hold that terms like *thing*, *object*, *physical object*, *space-occupier*, *entity*, and so on, are not sortals, despite their count-noun syntax, since they don't provide identity conditions (e.g., Hirsch, 1982, p. 38; Wiggins, 1980, p. 63; 1997, p. 418). Just as we can't count the black stuff that constitutes a black table, we can't count the physical objects that constitute it; the number could again be one (the table), five (the legs and top), six (the legs, top, and the table), and so on.

One way to square sortals with Spelke's physical objects is to note that Spelke's object concept is more specific than the ordinary notion of a physical object. Many things that we single out as objects don't move independently and aren't spatially separated from their backgrounds (as Hirsch, 1997, and Wiggins, 1997, have pointed out). Trees, mountains, houses, fences, fire hydrants, and sidewalks, among many other things, are typically fixed in place and would fail to

trigger an object concept that is sensitive only to movement and spatial isolation. Similarly, nonmoving parts of larger wholes often qualify as objects in the everyday sense, but not in the sense of independently moving, spatially separated entities. We speak of legs of tables, fenders of cars, handles of mugs, organs of animals, and other parts as objects in their own right, despite the fact that they usually occupy a fixed position with respect to the relevant larger entity. A Spelke-type object concept can't pick out such objects, and for this reason, it seems best to regard this concept as corresponding to a kind of primitive or *proto-object* (sometimes called a *Spelke-object*). Could *proto-object* be a sortal? Because the parts of a table, for example, aren't proto-objects (since they usually don't move on their own), counting the proto-objects that constitute a table doesn't pose the problem that counting physical objects does (Carey & Xu, 1999; Xu, 1997). A table is a single proto-object. (For arguments against the idea that *proto-object* is a sortal, see Ayers, 1997; Hirsch, 1997; and Wiggins, 1997.) However, the idea that both *proto-object* and lower-level terms like *cup* simultaneously function as sortals still conflicts with strong sortal theories (e.g., Wiggins, 2001) in which only a single sortal captures all the identity conditions for a particular object. See also Note 4 of this chapter for further difficulties with the idea of multiple sortals for single objects.

7. Experiments following Xu and Carey (1996) have found cases in which infants younger than 10 months are able to perform correctly in simplified versions of the is-it-one-or-two task (e.g., Wilcox & Baillargeon, 1998; Xu & Baker, 2005). The exact age at which infants succeed at such tasks is not of central interest here; however, some of the explanations for this early success do bear on the question of what knowledge they draw on when they anticipate two versus one object. Carey and Xu (2001, p. 194) argue that "when spatiotemporal evidence does not favor one solution over another, infants can use featural differences for object individuation" (see also Xu, 2003a). Thus, in Xu and Carey's original (1996) task, spatiotemporal information from the moving objects (the fact that the elephant and cup fall on the same trajectory) overrides featural differences that would otherwise serve to distinguish the objects, causing errors for the younger infants. Older infants are able to marshal sortals that, in turn, overcome the misleading spatiotemporal facts. However, featural differences (e.g., shape and size changes) are precisely the kinds of properties that *don't* individuate objects, according to the philosophical theories of sortals described earlier (e.g., Strawson, 1959). To the extent that infants can use properties (without the support of underlying sortals) to distinguish the items in these experiments, the very difference between sortal and nonsortal predicates is placed in doubt (see Blok, Newman, & Rips, 2007, and Section 1.4.2 for further discussion).

8. Basic-level categories are sets like apples or chairs that are at a middle level of abstractness. They contrast with subordinate categories (such as Winesap

apples or Eames chairs), and superordinate categories (such as fruit or furniture). Rosch et al. (1976) provided evidence that basic-level categories possess advantages over subordinates and superordinates in a variety of cognitive tasks. Since Rosch et al.'s classic paper, investigators have raised questions about the stability of the basic level across tasks and amounts of expertise (see Murphy, 2002, for a discussion and defense of the basic-level notion). The present point, however, is simply that terms for basic-level categories tend to be those people favor in naming individual objects. Asked *What is it?* of a particular Winesap apple, people usually say *apple,* not *Winesap* or *fruit.*

9. We assume, along with Liittschwager and others, that proper names like *Jim* are rigid designators that always refer to the same individual across situations or possible worlds; see Kripke (1972). Participants who state that the transplant recipient is no longer Jim are therefore affirming that the recipient is no longer the same individual.

10. Criticism of the Closest Continuer theory has focused on this context sensitivity (e.g., Noonan, 1985; Williams, 1982). According to these criticisms, the question of whether x_0 is identical to x_1 cannot depend on the presence of individuals $x_2, x_3, \ldots$ that may also exist at the same time as x_1. The appeal of this idea (sometimes called the *only-x-and-y* principle) arises from the intuition that the identity of an individual is a relation between the individual and itself, and therefore cannot be affected by the presence of other things. But whether or not this is a correct metaphysical rule (Nozick, 1981, argues against it), considering alternatives seems an inevitable part of *recognizing* the identity of objects, which is the process in (1) that I hope to clarify. This context sensitivity is on a par with similar effects in judgments of similarity (e.g., Tversky, 1977) and choice (e.g., Shafir, Simonson, & Tversky, 1993).

11. Although Nozick's model blocks intransitivities of the sort just described, there is another way in which both Nozick's model and my own allow for intransitivities. Suppose object x_0 exists at time t_0, x_1 at t_1, and x_2 and x_2' at t_2. Then x_1 might be the closest continuer of x_0, and x_2 the closest continuer of x_1, but x_2' might be the closest continuer of x_0. In the experiments to be reported here, however, I consider only situations involving two time points; so no evidence exists on whether people produce this type of intransitivity.

12. This assumption is also factually correct. Although it might seem icebergs would have to grow before they can shrink, in fact icebergs are created when they break off from ice shelves in Arctic or Antarctic regions.

13. I thank Douglas Medin for suggesting this idea.

14. This experiment also contained a second part in which participants judged which of two icebergs later found in the same vicinity was Sample 94. Like

the two-choice condition in the first experiment, this was intended to test the quantitative version of the Causal Continuer theory. In general, the results were again favorable to the model; for details, see Rips, Blok, and Newman (2006).

15. It is possible to object that "causal integrity" itself presupposes sortal information, since what's integral in one domain may not be in another. But we are not taking causal integrity as the basic explanatory concept here. What *is* basic is the Causal Continuer model's evaluation of identity based on causal factors, and our use of "causal integrity" is meant as a stand-in for this evaluation. Since the model appears to account for identity judgments in domains as diverse as animals and icebergs, there is evidence that it applies successfully in a domain-general way. See Blok et al. (2005, 2007) for further discussion.

16. Psychological essentialists believe that, although people don't know the essential properties for a category, they nevertheless believe there are some (S. A. Gelman, 2003; Medin & Ortony, 1989). See Chapter 4 of this book for a discussion. But the same tactic will not work for psychological sortalists. In order to identify objects over time, it is usually not enough for people to believe that a category has some identity conditions or other (i.e., to have a placeholder for these conditions); they have to know exactly what the conditions are in order to identify the objects via principle (4).