Intention & Intentionality
Malle, Moses & Baldwin
MIT Press 2001

9

Making Sense of Human Behavior: Action Parsing and Intentional Inference

Jodie A. Baird and Dare A. Baldwin

Human behavior is sufficiently complex to render the task of interpretation formidable. Behavior tends to flow continuously, lacking pauses to identify boundaries between distinct actions. Despite this complexity, adults and even preschoolers readily make sense of human behavior, identifying actions within the behavior stream and interpreting them in terms of intentions and what they reflect about other mental states (Malle and Knobe 1997a; Schult and Wellman 1997). How is this achieved? We will explore the possibility that a basic ability to detect structure in human behavior supports the system of intentional understanding. In particular, we will suggest that there may be physical and temporal features of action that correlate with the initiation and completion of intentions. Sensitivity to this structure would enable observers to extract the portions of behavior that are meaningful for understanding an actor's intentions. Moreover, infants as well as adults may possess these structure-detection skills, laying the groundwork for the subsequent development of genuine intentional understanding.

In the first section, we present new research documenting adults' and infants' sensitivity to structure inherent in intentional action. We follow this with a discussion of the kinds of mechanisms that might subserve these structure-detection skills and to what extent such mechanisms might be domain-general or domain-specific.

Adults' Action Parsing

Previous research on action parsing has concentrated on adults' ability to identify units within the behavior stream when asked to do so deliberately. In Newtson's (1973, 1976) unit-marking procedure, for example, observers

watch a sequence of behavior (for example, a man repairing a motorcycle) and press a button on an event recorder when they determine that one meaningful action has ended and another has begun. This technique has revealed strong agreement among adult observers regarding action-unit boundaries: Their judgments reliably cluster within intervals of behavior ranging from 1 second to 5 seconds (Newtson 1976).

Despite consistent evidence of adults' capacity to detect units in the behavior stream, the question of precisely what defines these units remains controversial (see, e.g., Wyer and Srull 1989). Newtson and his colleagues have suggested that unit boundaries correspond with distinctive changes in the stimulus array, such as bodily shifts in the actor's position (Newtson and Engquist 1976; Newtson, Engquist, and Bois 1977). Avrahami and Kareev (1994) have argued instead that a sequence of behavior becomes perceptible as a unit only after it has been observed many times across varying contexts. However, neither of these proposals addresses how the units achieved in parsing relate to a conceptual-level interpretation of action. At the conceptual level, it seems clear that the structure derived from observing others in action tends to be framed in terms of ideas about the actors' goals and intentions. In explaining the actions of others, adults refer to psychological explanations, citing beliefs, desires, and intentions as the primary factors underlying human behavior (Heider 1958; Malle 1999, this volume). The central role of intentions in adults' folk theories of mind and behavior hints at the possibility that observers spontaneously segment the behavior stream into units that coincide with its intentional structure.

Both the problem we are outlining and our proposal for its solution are reminiscent of issues that have emerged in the domain of speech processing. Much like human behavior, speech proceeds as a continuous flow, lacking pauses to indicate appropriate segmentation. Fodor and Bever (1965) hypothesized that adults parse speech by perceiving it according to the constituent structure defined by underlying syntactic principles. To test this idea, Fodor and Bever (p. 415) capitalized on "the tendency of a perceptual unit to preserve its integrity by resisting interruptions." Adult participants in their study listened to tape recordings of sentences in which a "click" (a burst of white noise) was heard somewhere in each sentence. Their task was to recall the precise placement of the click within the sentence. Fodor and Bever reasoned that if grammatical constituents operate

as units in adults' processing of the speech stream, then participants' subjective placement of the clicks should cluster at the boundaries of such constituents. This is precisely what they found. When clicks occurred at major constituent boundaries, participants' judgments were quite accurate. In contrast, when clicks occurred elsewhere in the sentence, participants' judgments erred in the direction of major constituent boundaries. These findings support Fodor and Bever's suggestion that speech is processed with respect to its underlying constituent structure.

To test the analogous question in the action domain—whether observers' spontaneous organization of behavior coincides with an analysis of the actor's intentions—we applied a variant of Fodor and Bever's click methodology to the study of action parsing (Baird, Baldwin, and Malle 2000). We first videotaped everyday action sequences, such as a woman cleaning her kitchen. She washes a dish, hangs a fallen towel, puts ice cream in the freezer, and so on. The movement in these action sequences flows continuously, with no physical pauses to identify boundaries between intentions. We digitized the action sequences and then artificially inserted tones in two different kinds of locations: *endpoint* tones coincided with the completion points of intentional action, whereas *midpoint* tones interrupted the ongoing action just before the completion of intentions (figures 9.1 and 9.2). Our logic was as follows: If the boundaries between intentions coincide with psychologically salient portions of the behavior stream, then adults should demonstrate greater success at recalling the location of endpoint tones (which highlight these boundaries) than at recalling the location of midpoint tones (which bear no relation to these boundaries).

Rather than rely on our own intuitions to select the locations of endpoint and midpoint tones, we presented the action sequences to a group of adult coders. Their first task was to select the portions of action that were meaningful in terms of understanding the actor's intentions. Next, they were asked to identify the point at which each of these action units was completed; that is, the point at which the intention was fulfilled. We instructed the coders to identify intentions at two levels of analysis. At the first level of analysis, coders identified intentions such as hanging a towel and putting ice cream in the freezer. We called these *tasks*. At the second level of analysis, coders identified intentions such as picking up the ice cream and grasping the freezer door. We called these *smaller actions*. In our instructions, we described these

Figure 1
A frame in which an endpoint tone occurred.

Figure 2
A frame in which a midpoint tone occurred.

two levels to the coders using examples from a different, unrelated action scenario to avoid influencing their judgments. Coders' judgments turned out to show remarkable agreement. At both levels of analysis, coders tended to identify the same intentions, and their judgments of completion points were tightly clustered. On the basis of these judgments, we determined the differential placement of endpoint and midpoint tones in two kinds of stimulus action sequences: long action sequences (with tones at the task level) and short action sequences (with tones at the smaller-action level).

After constructing the stimuli, we presented a new group of adults with four different action sequences, two of each length. Each action sequence had two endpoint tones and two midpoint tones in alternating, counterbalanced order. To eliminate auditory interference with the presentation of the tones, action sequences were presented with their original soundtracks turned off. In the presentation phase of the procedure, adults viewed an action sequence once, with the instruction to pay close attention to the locations of the four tones. In the test phase, they watched the same action sequence again, this time without any tones, and were asked to click a computer mouse at the precise points at which they remembered the tones having previously occurred.

There are at least two reasons to predict that adults' subjective placement of endpoint tones should be more accurate than their placement of midpoint tones. First, tones located at endpoint locations correspond to boundaries between intentions. Thus, if intention boundaries coincide with psychologically salient junctures in behavior, adults should demonstrate greater accuracy in recalling endpoint locations relative to midpoint locations. Second, processing demands are lower at endpoint locations. When a tone occurs at an endpoint, adults should have completed their processing of the intention and thus should be able to concentrate solely on the task of processing the tone's location. When a tone occurs at a midpoint, however, adults should be in the midst of processing the intention in addition to the tone's location. Under these dual task conditions, their placement of midpoint tones should suffer relative to their placement of endpoint tones.

To examine the degree of error adults displayed in judging the locations of endpoint and midpoint tones, we first analyzed their absolute error scores in response to the two types of tones.[1] These scores were substantially smaller for endpoints (M = 430 milliseconds, SD = 173) than for midpoints (M = 643 milliseconds, SD = 261), confirming that adults were more

accurate in their placement of endpoint tones than in their placement of midpoint tones. To investigate the direction of adults' error (for example, anticipation vs. delay), we next calculated their signed error scores in response to endpoint and midpoint tones. On average, adults reported endpoint tones to have occurred a mere 7 milliseconds (SD = 306) after their actual locations. In contrast, they demonstrated significant delay in their placement of midpoints: Adults judged midpoint tones to have occurred, on average, 379 milliseconds (SD = 393) after they actually did. Hence, adults misremembered the midpoint tones as having occurred much closer to the actual endpoint locations. This considerable delay in the placement of midpoint tones supports the idea that intention boundaries define units in adults' parsing of the behavior stream.

A potential concern with our methodology is that factors other than the meaningful placement of endpoint and midpoint tones within the action sequence—for instance, the timing of the tones relative to each other rather than the relation between the tones and the action—may have driven the pattern of errors adults displayed. In fact, a similar issue was raised in response to Fodor and Bever's original study. Reber and Anderson (1970) were concerned that nonlinguistic factors could have been responsible for adults' placement of the clicks at constituent boundaries. To test this, Reber and Anderson conducted a study in which they replaced linguistic sentences with streams of broad-band white noise. The patterns of errors adults displayed in response to these meaningless messages were similar to the error patterns evidenced in response to grammatical sentences, suggesting that nonlinguistic mechanisms could have been responsible for Fodor and Bever's findings.

To control for a similar possibility in our own research, we conducted a follow-up study in which we replaced the human action sequences with sequences of colored images. These new sequences preserved the precise timing of the tones, yet lacked meaningful action information. Despite identical tone information, the predicted performance pattern demonstrated in response to the human action sequences disappeared in response to the color sequences, indicating that the systematic performance differences evident in the first study were due to the meaningful placement of endpoint and midpoint tones within the action sequence, and not to other factors such as the rhythmic timing of the tones. Taken together, the findings from these two studies suggest that adults do indeed parse human behavior along intention boundaries. Moreover, this research provides the first demonstration of a link between adults' parsing of the behavior stream and their conceptual interpretation of intentional action.

Infants' Action Parsing

The relation between adults' organization of behavior and their reasoning about behavior raises the possibility that action parsing abilities, if possessed by young infants, might critically subserve the ontogeny of genuine intentional understanding. To investigate whether infants possess skills for parsing action, we tested the ability of 10–11-month-olds to detect disruptions to the structure inherent in intentional action (Baldwin, Baird, Saylor, and Clark, in press). Using a variant of the habituation/dishabituation paradigm, we presented infants with digitized video sequences of everyday action; for example, one sequence depicted a woman reaching to grasp a towel and hang it on a rack. In a familiarization phase, infants viewed the same action sequence repeatedly across several trials. In a subsequent test phase, infants viewed two different versions of the original action sequence, this time with still-frame pauses inserted at certain points in the course of action. The *intention-completing* test version, like the endpoint tones in the adult work, highlighted the boundaries between intentions with a still-frame pause occurring just as the actor completed an intention. In contrast, the *intention-interrupting* test version disrupted the intentional structure of the action sequence with a still-frame pause occurring midstream as the actor pursued an intention. In both versions, the pauses suspended the ongoing action for 1.5 second without deleting any information. Moreover, in an effort to eliminate any auditory differences between the intention-completing and the intention-interrupting videos, the still-frame pauses interrupted the original soundtrack of the actor's movements and vocalizations equally often in the two versions.

Infants viewed four familiarization trials in which the unjunctured action sequence was presented. This familiarization phase gave infants adequate opportunity to go beyond superficial characteristics in their processing of the action information. They then viewed two trials each of the intention-completing and intention-interrupting test videos in alternating,

counterbalanced order. If infants parse behavior according to the structure inherent in intentional action, they should look longer at the intention-interrupting test videos (which violate this structure) than at the intention-completing test videos (which preserve this structure).

Infants' looking decreased from the first to the last familiarization trial, indicating that they indeed processed the unjunctured action sequence. When subsequently presented with the two types of test videos, infants demonstrated renewed interest—that is, longer looking times relative to the last familiarization trial—only in response to the intention-interrupting video, indicating that this video, but not the intention-completing video, violated their expectations. These findings demonstrate that infants, at least by 10–11 months, readily parse ongoing behavior along intention boundaries.

However, these results leave open the possibility that infants had a starting preference for the intention-interrupting test videos over the intention-completing test videos due to basic salience differences between the two types of test videos. To test for this possibility, we showed a new group of 10–11-month-olds just the intention-completing and intention-interrupting test videos without first familiarizing them with the unjunctured action sequence. Infants in this study looked equally long at the two types of test videos, clarifying that the longer looking at the intention-interrupting video in the first study was due to infants' sensitivity to violations of the structure of intentional action and not to any starting preference for that test video.

Possible Mechanisms

Our data provide clear evidence that infants as well as adults detect structure in action that coincides with the initiation and completion of intentions. At the same time, we cannot yet provide a complete account of *how* such structure is detected. At present, all our findings are consistent with both high-level and low-level explanations. On the high-level account, a conceptual understanding of the actor's intentions and goals may have driven structure detection in a top-down fashion. For example, knowledge of typical kitchen activity may have facilitated parsing of the action sequence. Alternatively, parsing may have been driven by low-level perceptual skills for recovering physical and temporal structure in the behavior stream that coincides with its intentional structure. For instance, as actors initiate intentions in the physical world, they first locate relevant objects with their sensors (typically the eyes), then move toward those objects, contact and manipulate them, and ultimately release contact as intentions are completed. This highly typical sequence is accompanied by a characteristic temporal dynamic: Movements in the sequence have a ballistic quality once initiated and follow one after another in rapid succession. A basic sensitivity to such predictable sequences of movement within the behavior stream could enable observers to extract just the right units of action for drawing inferences about the actor's intentions.

In the case of adults, we suspect that both mechanisms—a top-down understanding of the particular intentions involved and a bottom-up skill for identifying structure in action—likely operated to generate our findings. The action sequences adults viewed in these studies depicted well-known, everyday activities about which adults unquestionably possess sophisticated knowledge and expectations. This makes it likely that high-level mechanisms played some role in adults' processing of the action sequences. However, there are at least two theoretical reasons to suspect the additional involvement of low-level mechanisms. First, to the extent that there are indeed physical and temporal correlates of intentional action, it could only facilitate adults' action parsing to be sensitive to such structure. Second, a basic capacity to detect structure in action may be useful in the face of novel action sequences for which preset expectancies about the relations between actions and intentions will not suffice. Though it may in fact be rare for adults to observe extended novel action sequences, we suspect that there are novel elements in many of the action sequences that adults observe. When such novelty is encountered, adults must generate new action-intention relations. Doing so relies on an ability to organize the action itself in the absence of conceptual interpretation, and to use such action analysis as a basis for inferences about intentions.

In the case of infants, once again both the high-level and the low-level mechanisms postulated could account for the action-processing abilities recently documented in our own laboratory and in several others (Wellman and Phillips, this volume; Woodward, Sommerville, and Guajardo, this volume; see Povinelli, this volume, for a similar point). No existing data clarify this issue. It is quite possible that both high-level and low-level mechanisms are operative in infants' action processing. However, we

strongly suspect that a low-level mechanism plays an especially important role in the structure-detection ability infants have displayed in our own research. For one, infants younger than 10 or 11 months are known to be capable of sequence learning (Clohessy, Posner, and Rothbart 1992; Haith, Hazan, and Goodman 1988) and statistical covariation detection (Aslin, Saffran, and Newport 1998; Saffran, Aslin, and Newport 1996), skills akin to those we are proposing here. Moreover, a low-level skill for detecting meaningful structure in the behavior stream seems to be a prerequisite for infants' emerging sophistication in the realm of intentional understanding. Each day, infants confront a host of novel action sequences for which they lack the world knowledge needed to understand the relevant intentions. A basic sensitivity to physical and temporal regularities in the behavior stream that coincide with intentions would enable infants to identify the meaningful portions of an action sequence even when they do not yet understand the particular intentions involved. Armed with the appropriate units, infants would then be well positioned to engage in further processing, ultimately yielding the discovery of those intentions. Thus, a low-level mechanism of the kind we are proposing would be a critical prerequisite for the developmental emergence of genuine intentional understanding.

Intentional Understanding: A Two-Tier System

Intentional understanding is the ability to draw inferences about psychological motivations guiding human action. In our view, at least two mechanisms are crucially involved in the development of intentional understanding. First, as we have already suggested, a low-level mechanism for detecting statistical structure in human action enables observers to identify relevant units in the behavior stream. For the achievement of genuine intentional understanding, however, a higher-level mechanism is still required to make sense of the resulting units in terms of psychological motivations. In what follows, we will address some basic questions one might ask regarding the nature and functioning of this two-tier system.

Are These Systems General or Specific?

Both of the mechanisms we are proposing could well operate in a domain-general fashion, supporting structure detection and inferential reasoning in

any arena with similar processing demands. With respect to the low-level mechanism, the language domain is the arena in which infants' structure-detection skills have been documented most clearly. Specifically, the detection of statistical regularities seems to enable infants to extract word-like units within the speech stream (Aslin et al. 1998; Saffran et al. 1996). This kind of statistical recovery has been modeled successfully by associative networks, clarifying the low-level nature of these structure-detection skills (McClelland and Plaut 1999). As we have suggested, low-level structure-detection skills of this same kind may have generated infants' parsing of intentional human action in our studies. Moreover, these skills could help infants to recognize important structural *differences* among various forms of movement in the physical world. For example, infants need to acquire the ability to distinguish between intentional action and inanimate motion. Furthermore, intentional action must be distinguished from behavior that is accidental or inadvertent. As it turns out, there is evidence suggesting that young infants appreciate some basic differences between human action and inanimate motion, although the mechanism underlying this ability is not yet clear. For example, 7-month-olds recognize that humans, but not inanimate blocks, can cause one another to move in the absence of direct physical contact (Woodward, Phillips, and Spelke 1993). Infants might even be able to make the important distinction between intentional action and unintentional behavior on a structural basis. As an illustration of the structural differences between these two classes of human movement, consider a clumsy mistake such as dropping a towel while reaching to place it on a rack. In this unintended scenario, the initial trajectory of motion involving the towel is interrupted; additionally, eye gaze is directed toward the floor belatedly, following (rather than preceding) the towel's fall.

Regarding the high-level mechanism, we believe inductive reasoning plays a crucial role across a variety of domains in assigning functional significance to units extracted via low-level structure detection. In the language domain, for example, a low-level mechanism may enable the detection of word-like units within the speech stream, but an inferential mechanism is necessary to assign semantic content to those units once identified. Likewise, units of action extracted by means of structure detection are not conceptually meaningful until the intentions motivating them are inferred. Explicit models of domain-general inductive mechanisms such as those we

are proposing have been constructed to account for a range of relevant phenomena in adult reasoning; see, e.g., Gentner and Markman 1997 and Holland, Holyoak, Nisbett, and Thagard 1986.

In suggesting that intentional understanding may be driven at least in part by domain-general mechanisms, our proposal is notably different from others currently available in the literature. For example, both Premack (1990) and Baron-Cohen (1995) credit infants with biologically prepared, fully specified "intention detection" systems. In Premack's account, this system not only detects that an intention is occurring but also recognizes the content of at least some specific intentions (Premack and Premack 1995). In contrast, we are proposing that early in infancy *structure* can be detected in action, not necessarily the content of intentions per se. An interpretation of action in genuinely intentional terms requires the operation of a higher-level inferential mechanism.

How Do These Systems Interact?

In adults, we suspect that these two mechanisms—low-level structure detection and high-level inferencing—work in close concert, operating in a highly interactive and parallel fashion to guide people's rapid processing and interpretation of others' action. For example, when adults encounter novel actions in a foreign culture, the low-level mechanism operates to derive meaningful portions of action while the high-level mechanism simultaneously strives to interpret those units in intentional terms.

Both high-level and low-level mechanisms are likely operative in infancy. However, unlike adults, young infants presumably know little of the diverse objects cluttering our physical world. Such knowledge plays a central role in driving appropriate, high-level inferences about intentions in any given context. For example, to fully understand intentions such as those underlying the kitchen-cleanup scenario we presented, one must grasp the relevance of water to dirty dishes, that of freezers to melting ice cream, and indeed that of a clean kitchen to an orderly life. Without such world knowledge, infants' processing of action cannot be driven by the kinds of specific inferences about intentions that likely aid adults' action processing. Hence the benefit of a low-level structure-detection mechanism.

A low-level mechanism of the kind we are proposing would provide infants with an initial organization of the behavior stream in the absence of

sophisticated understanding of specific intentions. Specifically, low-level structure-detection abilities yield relevant units of behavior from which the high-level inferential mechanism can begin to generate intentional inferences. We are not denying that infants possess domain-general inferential abilities; in fact, there is evidence to suggest that they do by at least 9 months (Baldwin, Markman, and Melartin 1993). We are simply arguing that in order for their inferential skills to produce the appropriate inferences about intentions in the domain of action interpretation, the low-level mechanism must first provide the relevant units on which to base such inferences. Thus, the low-level mechanism likely plays a more crucial role for infants than adults.

Questions for the Future and Concluding Remarks

Our two-tier account of the system for intentional understanding raises a number of questions for future investigation. First, with respect to the low-level mechanism, precisely what characterizes the structure of intentional action? It will be important to identify, for example, the physical and temporal properties of action that covary with transitions between intentions. Also, in what ways does the structure of intentional action differ from the structure inherent in other forms of physical movement, such as unintentional behavior and inanimate motion? When observing any kind of movement, are adults, and perhaps even infants, biased to extract intention-relevant structure in preference to other kinds of available structure? Are there individuals with specific deficits in detecting structure within action?

Regarding the high-level inferencing mechanism, how do observers resolve which intention is relevant on any given occasion? Each and every action can be interpreted in many ways; how do observers constrain the field of possibilities to an appropriate set of specific intentions? Finally, who is capable of drawing inferences about intentions? Perhaps some individuals suffer disruptions in this ability. Sabbagh (1999) has recently argued that those with autism and right-hemisphere damage have impairments that specifically undercut their ability to draw inferences about intentions. Moreover, there may be important species differences in inferential reasoning that yield corresponding differences in intentional understanding.

Along these lines, Povinelli (this volume) argues convincingly that chimpanzees and other higher primates are skilled at detecting structure in action but lack the inferential ability necessary to interpret this structure in genuinely intentional terms. If so, then divining the intentional significance of one another's actions may be a uniquely human preoccupation. We join with Povinelli in hoping to focus future inquiry on the precise nature of the relation—in both evolution and development—between skills for low-level behavior analysis and inferences about intentions.

Such questions aside, our findings clarify that infants as well as adults readily detect structure in action that is relevant for drawing inferences about intentions. In our view, low-level action analysis skills likely enable such structure detection. These basic parsing abilities gain infants access to appropriate units within complex, continuous behavior, and help adults similarly process novel action sequences for which they lack a ready motivational explanation. In this sense, structure-detection skills play a crucial role in potentiating our lifelong pursuit of interpersonal understanding.

Acknowledgments

This work was supported by a National Science Foundation Graduate Research Fellowship to the first author and by a NSF New Young Investigator award (No. 9458339) and a John Merck Scholars Award to the second author. Our thanks to Janet Astington, Rebecca Brand, Diego Fernandez-Duque, Bertram Malle, Lou Moses, Daniel Povinelli, and Mark Sabbagh for helpful comments.

Note

1. The same pattern of results was found for both long and short action sequences. For clarity, only the findings from the short action sequences are presented here.

10

Desire, Intention, and the Simulation Theory

Alvin I. Goldman

A common goal of philosophy of mind and psychology is to understand how people acquire, represent, and deploy such mentalistic concepts as desire, intention, and belief. Developmental psychologists call this subject *theory of mind*. Although this label may not commit them to the so-called *theory-theory*, almost all developmentalists (with the major exception of Paul Harris) in fact favor the theory-theory approach to the subject. A greater division of opinion prevails among philosophers, with a dedicated coterie advocating a competing approach known as the *simulation theory* (Gordon 1986; Heal 1986; Goldman 1989; Currie 1995). In this chapter I argue that the vaunted theory-theory (TT) is not on very secure ground, that it suffers from serious weaknesses not adequately appreciated by many of its advocates, and that the simulation theory (ST)—at least, its "introspectionist" variant—promises to be a cure for many of these ills. I discuss these issues with special attention to desires and intentions.

Theory-Theory and Conceptual Change

The standard version of TT is a combination of two theses: (1) that common-sense psychological concepts are theoretical concepts, similar in all relevant respects to the theoretical concepts of physical science, and (2) that people detect psychological states in themselves and in others by making theoretical inferences. In trying to defend these two theses, developmentalists typically appeal to multiple stages in children's grasp of mentalistic concepts. First, they find evidence of performance changes in mental-state attributions. Second, they interpret these performance changes as changes in children's conceptual repertoire concerning mentalistic states. Third, they