1 Introduction

1.1 Motivation

Our experience of everyday social life is deeply shaped by the actions that we see others perform: consider a parent carefully watching her infant try to feed herself, a fan watching a tennis match, or a pottery student observing her teacher throw a pot. Although we may sometimes pause momentarily in puzzlement (what is my neighbour doing up there on his roof?) or be caught by surprise (by a partner's sudden romantic gesture), we normally understand others' actions quickly and without a feeling of expending much effort. By doing so, we unlock answers to important questions about the world around us: What will happen next? How could I learn to do that? How should I behave in a similar situation? What are those people like?

How, then, do we understand observed actions? The simplicity of this question, and the fluency of action understanding, obscures the complexity of the underlying mental and neural processes. To start to answer it, and in contrast to several recent valuable perspectives (e.g. Kilner, 2011; Oosterhof et al., 2013; Pitcher & Ungerleider, 2021; Tarhan & Konkle, 2020; Thompson et al., 2019; Tucciarelli et al., 2019; Wurm & Caramazza, 2021) we do not focus first on possible brain mechanisms (including the possible role of mirror neurons; see Bonini et al., 2022; Heyes & Catmur, 2022). Instead, first thinking about the problem in terms of Marr's (1982) *computational* level, we ask: why would an observer attend to the actions of others? A reasonable answer to this question might be: Observers attend others' actions to learn about the meaning and outcomes of different action kinds; to establish causal links between actors' actions and their goals, states, traits, and beliefs; and to use that learned knowledge to make predictions about the social and physical environment, and to extend one's own action repertoire. (Although beyond the focus of this review, we also sometimes attend others' actions for pure enjoyment, e.g. when watching ballet or figure skating; e.g. Christensen & Calvo-Merino, 2013; Orgs et al., 2013). Achieving these multiple complex aims requires suitable mental representations and processes – *algorithms* in Marr's (1982) terms. That is the main focus of Section 2 of this article. In Section 3, we go on to describe key neuroscientific evidence on action understanding (focusing on Marr's *implementation* level), drawing links to the concepts and constructs described in Section 2. In the final section, we identify directions for future research that are highlighted by this review.

1.2 Definitions and Scope

A survey of the literature in neuroscience, psychology, computer science, and cognitive science reveals a proliferation of terminology (action recognition, action comprehension, action identification, action perception, action observation, action interpretation, activity recognition) and equally diverse definitions. These make related, but not always consistent, assumptions and distinctions (Table 1) that may in part be due to different aspects of action that are highlighted in different experimental paradigms (Figure 1). This diversity is to be expected, given the complexity of the topic and the need to simplify it to gain traction. For this review, we adopt the term 'action understanding' as an umbrella term of convenience, to refer in general to the act of making sense of viewed human actions, and we avoid making further terminological distinctions. We resist the temptation to provide a single, concise definition of action understanding, preferring that this should emerge from the breadth of behaviours, cognitive mechanisms, and brain systems that we describe. However, some basic assumptions provide a grounding: we are concerned with observable behaviours that are intended to effect changes in the physical world or on others' minds.

What topics fall under the broad umbrella of 'action understanding'? We focus here on human action understanding, so we do not consider purely engineering-led approaches such as AI systems for what is typically known as

Table 1 Definitions of action understanding.

Gallese et al. (1996)	*'the capacity to recognize that an individual is performing an action, to differentiate this action from others analogous to it, and to use this information in order to act appropriately'*
Rizzolatti et al. (2001)	*'We understand actions when we map the visual representation of the observed action onto our* ***motor*** *representation of the same action'.*
Kohler et al. (2002), *Science*	Audio-visual mirror neurons might contribute to action understanding by evoking 'motor ideas'
Fogassi et al. (2005), *Science*	Mirror neurons selectively encode the goals of motor acts and thus facilitate action understanding
Bonini & Ferrari (2011)	*Action recognition: 'know again, recall to mind'; the ability to form a link between sensorimotor description and motor representations*
Rizzolatti & Sinigaglia (2016)	*'the outcome to which the action is directed'*

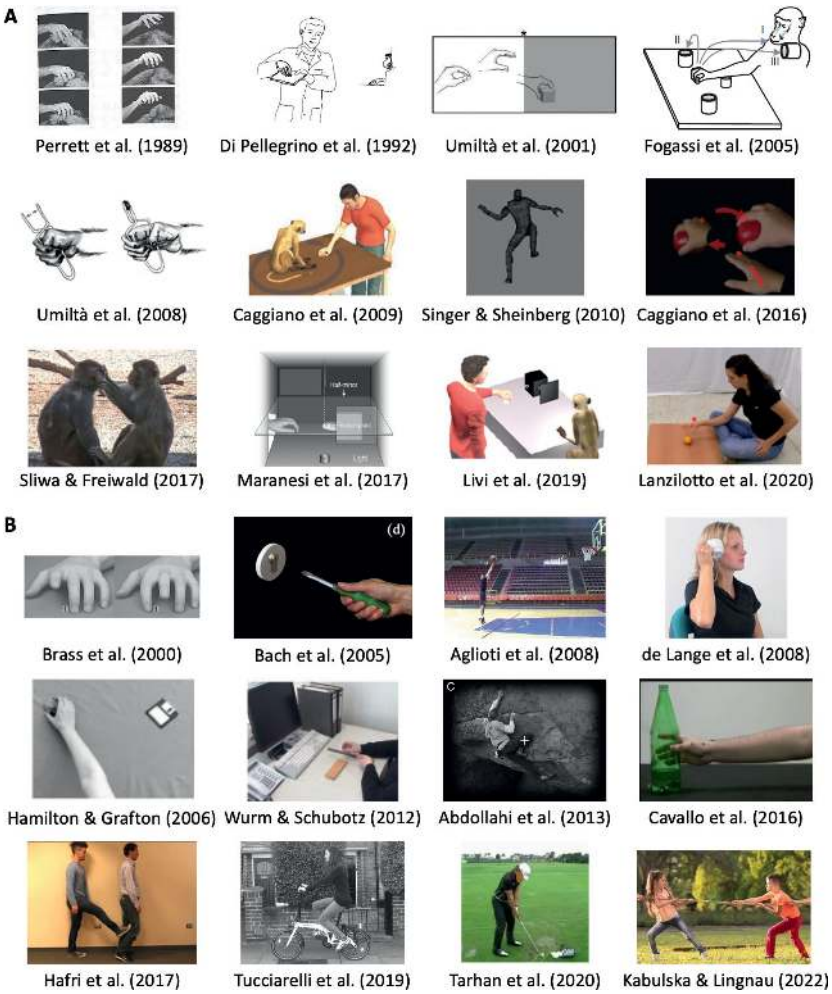

Figure 1 What we talk about when we talk about action understanding. **A:** Examples of paradigms used in the monkey literature. **B:** Examples of paradigms used in the human literature. These examples give a sense of the wide variety of stimuli and tasks used in this literature, which may include schematics, still images, animations, or movies of typical or atypical manual or whole-body actions, either in a natural or a constrained context. The diversity of these examples is matched by the diversity of terminology and definitions adopted in the action understanding literature (see Table 1).

action classification or activity recognition in that literature (Muhammad et al., 2021; Vrigkas et al., 2015). As vision is at the heart of most treatments of human action understanding, we focus on understanding seen real-world actions

(but see Camponogara et al., 2017; Repp & Knoblich, 2004, for discussion of action understanding in other modalities). Evidence from animals is reviewed for its influences on thinking about human action understanding. We set aside the interpretation of actions and interactions that are conveyed symbolically, such as the decisions of a partner in an economic game like the Prisoners' Dilemma (e.g. Axelrod, 1980). Finally, we focus on understanding by typical healthy adult observers in exclusion of neuropsychological or neuropsychiatric populations. The logic for this is that while action understanding difficulties are associated with (for example) autism, schizophrenia, or semantic dementia, it is not clear that this is necessarily a central feature of those conditions (see e.g. Cappa et al., 1998; Cusack et al., 2015; Frith & Done, 1988). Action clearly is central to apraxia, however in that case definitions and diagnostics tend to focus on patients' *production* of appropriate gestures and skilled actions, particularly those relevant to tool use (Baumard & Le Gall, 2021) rather than understanding per se (but see e.g. Kalénine et al., 2010). That said, these difficulties may be informative for our thinking about the different computations and algorithms involved in action understanding; the same caveat applies to developmental evidence (Reddy & Uithol, 2016; Southgate, 2013).

Other, more specific action-related topics have recently been reviewed elsewhere: these include the perception of social interactions (McMahon & Isik, 2023; Papeo, 2020; Quadflieg & Westmoreland, 2019), the execution of joint or collaborative actions (Azaad et al., 2021; Sebanz & Knoblich, 2021), and visual perception of biological motion, especially from 'point-light' displays (Blake & Shiffrar, 2007; Thompson & Parasuraman, 2012; Troje & Basbaum, 2008).

1.3 General Principles

Two principles that have motivated many researchers' thinking about action understanding recur in our review. First, inspired by theories of hierarchies in the motor system (Georgopoulos, 1990; Harpaz et al., 2014; Turella et al., 2020; Uithol et al., 2012), actions are often described at different hierarchical levels (see Table 2). These include *kinematics* (the *how* of an action), the action *kind* (the *what* of an action), and the *intention* (the *why* of an action). These levels have strong implications for the representations and processes that are required for action understanding, and accordingly we adopt this three-way distinction to structure Section 2. The idea that actions can be described at multiple levels implies that action understanding may emphasize one of these levels over the others, depending on the observer's goals. For example, a basketball player who aims to improve his three-pointer performance might attend to the kinematics of the throw (e.g. the angle of the arm and the hand, the trajectory of the ball).

Table 2 Action understanding at different hierarchical levels.

Vallacher & Wegner (1989)	Actions can be identified on a range of different levels, from low level (how is the action performed?) to high level (why or with what is the action performed?)
Hamilton & Grafton (2007)	Muscle level (pattern of activity in all involved muscles) Kinematic level (shape of the hand, movement of the arm) Goal level (intention and outcome)
Spunt et al. (2011)	How vs What vs Why
Kilner (2011)	Kinematic level (trajectory and velocity profile) Motor level (processing and pattern of muscle activity) Goal level (immediate purpose of the action) Intention level (overall reason)
Wurm & Lingnau (2015)	Abstract level (generalization across different exemplars) Concrete level (exemplar-specific)
Thompson et al. (2019)	Action identification (e.g. precision versus whole hand) Goal identification (e.g. to grasp the cup) Intention identification (e.g. to quench thirst)
Zhuang & Lingnau (2022)	Taxonomic levels (superordinate, basic and subordinate level)

In contrast, a basketball player that aims to prevent a three-pointer by another player might attend to the intention of his opponent (e.g. by focusing on the gaze direction of the other player). This view conflicts with descriptions of action understanding as 'automatic', which would imply a process that unfolds independently of observer goals and the demands of other concurrent tasks that may 'load' cognition or perception. So in Section 2, we also describe different conceptions of automaticity and how they might play out in different action understanding situations.

Second, like any form of perception (Bar et al., 2006; De Lange et al., 2018; Hutchinson & Barrett, 2019; Rao & Ballard, 1999), action understanding enables *predictions* about what is likely to follow next, over timescales from seconds to years (Kilner et al., 2004; Oztop et al., 2005; Schultz & Frith, 2022; Umiltà et al., 2001). In some situations predictions are implicit (e.g. watching our tennis partner prepare to serve, a hunch that she will fault), and explicit in others (e.g. anticipating that the opposing tennis player will try to play a cross ball while one finds oneself in the opposite corner of the court). Predictions emerge across the hierarchical levels identified above. For example, from local cues such as hand or arm kinematics, gaze direction, and grasp preshaping, an

observer can make spatially and temporally precise predictions about how an action will unfold (McDonough et al., 2019), and about the target of a reaching movement or the intended use of a grasped object (Ambrosini et al., 2011, 2015; Amoruso & Finisguierra, 2019; Amoruso & Urgesi, 2016). At the same time, our semantic knowledge about different kinds of actions includes descriptions of their typical aims, and of the kinds of events that typically tend to follow (cf. Schank & Abelson, 1977). For example, observing a friend hand-washing the dishes implies that next they will be dried and put away. Finally, observing an action supports inferences about an actor's underlying goals and beliefs, enabling predictions about what future actions would be consistent with those beliefs, or further those goals, and indeed how that actor might behave in new situations even into the distant future.

2 What, How, and Why?

2.1 'What': Two Conceptions of Action Categorization

To answer the question 'what kind of action am I seeing now?' requires extracting visual information about the surrounding scene, the actors and their movements, objects, and the relationships among those elements. This perceptual evidence must be compared to stored representations of the actions that the observer knows about. The studies considered in this section have addressed two main research questions posed by those requirements: How is long-term knowledge about action kinds organized? And how is perceptual data matched to that knowledge?

Classifying an action requires the ability to generalize over variation caused by different viewpoints, lighting effects, occlusion, and other visual variables, just as in visual object recognition (see also Perrett et al., 1989). Further, a given action (e.g. chopping vegetables) may be carried out by many possible actors, using many possible objects, in many possible locations. That problem of *generalization* is complemented by the problem of *specificity,* which requires correctly excluding from a category exemplars that do not belong. Taking an analogy from objects, for example, one must understand that a robin (canonical exemplar) and a penguin (unusual exemplar) are both birds, but that a bat, despite numerous shared features with the bird category, is not. Figure 2 illustrates that similar problems arise for action understanding, where the challenge is to correctly include visually diverse exemplars while excluding attractive foils.

Finally (and also like objects), actions are well described by taxonomies that include an abstract (or 'superordinate') level, a basic level, and a subordinate level (Rosch et al., 1976; Zhuang & Lingnau, 2022). For example, 'playing tennis' may describe an action at the basic level that is part of a superordinate

Figure 2 Successful action understanding requires generalizing over highly distinct exemplars (e.g. of <chopping vegetables>; right side) including unusual ones (centre bottom image) while excluding highly similar non-exemplars (e.g. carving; left side).

category 'sporting activities' and also includes the subordinate level 'performing a forehand volley'. The basic level has been proposed to play a key role in object categorization, e.g. as evidenced by the number of features used to describe objects, and the speed of processing (Rosch et al., 1976). Zhuang & Lingnau (2022) recently reported similar results for actions. Specifically, participants produced the highest number of features to describe actions at the basic level (see also Morris & Murphy, 1990; Rifkin, 1985). Moreover, they verified action categories faster and more accurately at the basic and the subordinate level in comparison to the superordinate level. These findings suggest that the taxonomical levels of description proposed for objects have a homology in the long-term representation of action knowledge.

Action Spaces

One major approach to understanding the representation of action knowledge was influenced by previous work investigating the mental representation of objects (e.g. Beymer & Poggio, 1996; Edelman, 1998; Gärdenfors, 2004; Kriegeskorte et al., 2008a, b). These studies develop the idea that known actions are described by multidimensional 'spaces' (see Figure 3), in which each type of action occupies a point in that space (Dima et al., 2022; Kabulska & Lingnau, 2022; Lingnau & Downing, 2015; Thornton & Tamir, 2022; Tucciarelli et al., 2019; Watson & Buxbaum, 2014; Zhuang & Lingnau, 2022). Traversing along one hypothetical dimension, actions should vary systematically on one action