Abstract The concept of *agency* is important in philosophy, cognitive science, and artificial intelligence. Our aim in this paper is to highlight some of the issues that arise when considering the concept of agency across these disciplines. We discuss two different views of agency: agents as actors (the originators of purposeful deliberate action); and agents as intentional systems (systems to which we attribute mental states such as beliefs and desires). We focus in particular on the view of agents as intentional systems, and discuss Baron-Cohen's model of the human intentional system. We conclude by discussing what these different views tell us with respect to the goal of *constructing* artificial autonomous agents.

Key words: agency, intentional systems, logic, artificial intelligence

Them and Us*: Autonomous agents *in vivo* and *in silico*

Peter Millican and Michael Wooldridge

1 Introduction

As we look around our world and try to make sense of what we see, it seems that we naturally make a distinction between entities that in this paper we will call "agents", and other objects. An agent in the sense of this paper is something that seems to have a similar status to us as a self-determining actor. When a child deliberates over which chocolate to choose from a selection, and carefully picks one, we perceive agency: there is choice, and deliberate, purposeful, autonomous action. In contrast, when a plant grows from underneath a rock, and over time pushes the rock to one side, we see no agency: there is action, of a kind, but we perceive neither deliberation nor purpose in the action.

The concept of agency is important for philosophers (who are interested in understanding what it means to be a self-determining being) and for cognitive scientists and psychologists (who seek to understand, for example, how some people can come to lack some of the attributes that we associate with fully realised autonomous agents, and how to prevent and treat such conditions). However, the concept of agency is also important for researchers in computer science and artificial intelligence, who wish to build computer systems that are capable of purposeful autonomous action (either individually or in coordination with each other). If such artificial agents are to interact with people, then it must be helpful also to understand how people make sense of agency.

The aim of this paper is to survey and critically analyse various ways of conceptualising agents, and to propose what we consider to be a promising approach. Our discussion encompasses contributions from the literature on philosophy, cogni-

Peter Millican
Hertford College, Oxford OX1 3BW, UK e-mail: peter.millican@hertford.ox.ac.uk

Michael Wooldridge
Dept of Computer Science, University of Oxford, Oxford OX1 3QD, UK e-mail: mjw@cs.ox.ac.uk

* With apologies to Will Hutton.

tive science, and artificial intelligence. We start by examining two different views of agency:

- *First-personal view*. From this perspective, agents are purposeful originators of deliberate action, moved by conscious purposes.
- *Third-personal view*. From this perspective, agents are entities whose behaviour can be predicted and explained through the attribution to them of beliefs, desires, and rational choice.

Cutting across these perspectives is the issue of *higher-order intentional reasoning*, by which an agent may adopt the third-personal view of other agents and adapt its behaviour accordingly, based in part on the intentional states that it attributes to those other agents. We shall see some evidence that such reasoning — a distinctive characteristic of human beings in social groups — provides a plausible evolutionary driver of our own brain size and conspicuous "intelligence". Following a discussion of the human intentional system and the condition of autism (drawing on work by Simon Baron-Cohen), we turn to the question of agency *in silico*, and ask what lessons can be learned with regard to the construction of artificial autonomous agents.

2 Agency from the First-Personal Perspective

We will begin with the idea that *agents are the conscious originators of purposeful deliberate action*. As conscious beings ourselves, we naturally find this a compelling viewpoint, and it has understandably spawned many centuries of discussion about such thorny problems as free will, personal identity, and the relation between mind and body. Even if we leave these old chestnuts aside, however, the view raises other difficulties, which it will be useful to rehearse briefly.

First, there is the basic problem of how actions should be counted and individuated (which also arises, though perhaps less severely, from the third-personal perspective). Consider the following classic example, due to John Searle [31]. On 28 June 1914, the 19-year-old Yugoslav Nationalist Gavrilo Princip assassinated Archduke Franz Ferdinand of Austria, and thereby set in motion a chain of events that are generally accepted to have led to World War I, and hence the deaths of millions of people. This is, surely, one of the most famous deliberate actions in history. But if we try to isolate exactly *what action it was* that Princip carried out, we run into difficulties, with many different possibilities, including:

- Gavrilo squeezed his finger;
- Gavrilo pulled the trigger;
- Gavrilo fired a gun;
- Gavrilo assassinated Archduke Ferdinand;
- Gavrilo struck a blow against Austria;
- Gavrilo started World War I.

All six of these seem to be legitimate descriptions of what it was that Princip did, yet we are naturally reluctant to say that he simultaneously performed a host of

actions through the simple squeezing of his finger. We would like to isolate some *privileged* description, but can be pulled in different directions when we attempt to do so. One tempting thought here is that the remote effects of what Princip did are surely no part of his action: allowing them to be so would mean that people are routinely completing actions long after they have died (as well as performing countless actions simultaneously, e.g., moving towards lots of different objects as we walk). This line of thought naturally leads us to identify the *genuine* action as the initiation of the entire causal process in Princip's own body — the part over which he exercised direct control in squeezing his finger. But if we go that far, should we not go further? Princip's finger movement was caused by his muscles contracting, which was in turn caused by some neurons firing, which was caused by some chemical reactions... and so on. We seem to need some notion of basic or primitive action to halt this regress, but if such primitive actions are at the level of neuronal activity, then they are clearly not directly conscious or introspectible. This, however, makes them very doubtful paradigms of deliberate action, especially from the first-personal perspective whose focus is precisely on consciousness, and is therefore quite oblivious of the detailed activity of our muscles and neurons.

(As an aside, notice that when we consider the notion of agency in the context of computers, this threat of regress is, to some extent at least, mitigated. Computer processors are *designed* using an explicit notion of atomic action — in the form of an "atomic program instruction" — an indivisible instruction carried out by the processor.)

In reaction to these difficulties, a quite different tempting thought is precisely to appeal to our first-person experience, and to identify the *genuine* action with the effect that we consciously intend. But here we can face the problems of both too much, and too little, consciousness. For on the one hand, Princip plausibly *intended* at least four of the six action descriptions listed above, and again, this route will lead to posthumous action (since people routinely act with the conscious intention of bringing about effects after their death, such as providing for their children — see [20, pp.68-73] for the more general problem of trying to pin down the timing of extended actions). On the other hand, a great many of the actions that we perform intentionally are done without explicit consciousness of them, and the more expert we become at a skill (such as driving, riding a bike, typing, or playing the piano), the more likely we are to perform the actions that it involves with minimal consciousness of what we are doing (and indeed trying to concentrate on what we are doing is quite likely to disrupt our performance). Even when we do become fully conscious of acting in such a context — for example, when I suddenly swerve away on seeing a pedestrian fall into the road just ahead of my car — such activity is likely to *precede* our consciousness of it, and its emergency, "instinctive" nature anyway makes it an unlikely paradigm of conscious deliberate action.

In the face of these sorts of difficulties, many philosophers (notably Michael E. Bratman [6]) have come to prefer an account of intentional action in terms of *plans*. Here, for example, is the first approximate formulation by Mele and Moser:

> A person, *S*, intentionally performs an action, *A*, at a time, *t*, only if at *t*, *S* has an action plan, *P*, that includes, or at least can suitably guide, her *A*-ing. [25, p.43]

They go on to add further conditions, requiring that *S* have an *intention* which includes action plan *P*, and also that *S* "suitably follows her intention-embedded plan *P* in *A*-ing" [25, p.52] (for present purposes we can ignore here the additional conditions that Mele and Moser formulate to capture plausible constraints on evidence, skill, reliability, and luck). But importantly, intentionality is consistent with *S*'s having "an intention that encompasses, ... subconsciously, a plan that guides her *A*-ing" [25, p.45]. Seeing actions as falling into a pattern guided by a plan thus enables habitual or automatic actions to be brought into the account, whether they are conscious or not.

All this somewhat undermines the all-too-natural assumption that the first-personal point of view is specially privileged when it comes to the identification of, and understanding of, action. And as we shall see later, such theories of human action (e.g., Bratman's) have already borne fruit in work towards the design of practical reasoning computer agents. But in fact there is nothing here that precludes the idea that consciousness of what we are doing — and conscious reflection on it — plays a major role in *human* life and experience. A cognitive model that explains action in informational terms is perfectly compatible with the supposition that certain aspects of its operation may be available in some way to consciousness. For example, Goldman [19] sketches the model of action proposed by Norman and Shallice [28] and explains how conscious awareness "of the selection of an action schema, or a 'command' to the motor system" could play a role within it.

There might well, however, seem a threat here to our conception of human *free will*, if consciousness of what we are doing is seen as post-hoc monitoring of unconscious cognitive processes that have already taken place by the time we become aware of them. Such worries may be sharpened by recent research in neuropsychology, in which observations using MRI scanners indicated that the mental sensation of conscious decision can lag quite some time behind certain identifiable physiological conditions that are strongly correlated with the decision ultimately made. Experiments carried out at the Max Planck Institute for Human Cognitive and Brain Sciences in Germany suggested that it was possible to detect that a person had already made a choice, and what that choice was, up to *ten seconds* before the person in question was consciously aware of it [34]. Interpretation of such results is highly controversial, and there is clearly more work to be done in this area. We have no space to explore the issues here, but would end with four brief comments. First, we see no significant conflict between the idea that our thought is determined by unconscious "subcognitive" processes and the claim that we are *genuinely* free. To confine ourselves to just one point from the familiar "compatibilist" arsenal of arguments, the term "free choice" is one that we learn in ordinary life, and it would be perverse to deny that paradigm cases of such choice (such as a child's choosing a chocolate, with which we started) are genuinely free — if *these* aren't cases of free choice, then we lose all purchase on the intended meaning of the term. Secondly, it is entirely unsurprising that our conscious thinking should be found to correlate strongly with certain events in the brain, and such correlation does not imply that "we" are not really in control. On the contrary, neural processes are apparently the mechanism *by which* we reason and make choices; that they determine our thoughts

no more implies that "we" are not really thinking those thoughts than the transfer of visual signals along the optic nerve implies that "we" are not really seeing things (or, for that matter, that the electronic activity of its components implies that a computer is not really calculating things). Thirdly, we would resist any suggestion that the neurophysiological evidence points towards *epiphenomenalism* — the theory according to which mind and mental states are caused by physical (brain and body) processes, but are themselves causally inert (crudely but vividly, this takes the conscious mind to be a passenger in the body, under the illusion that it is a driver). If evolution has made us conscious of what we do, then it is overwhelmingly likely that this has some causal payoff, for otherwise it would be an outrageous fluke — utterly inexplicable in terms of evolutionary benefit or selection pressure — that our consciousness (e.g., of pains, or sweet tastes) should correlate so well with bodily events [27, §5]. Finally, there could indeed be some conflict between our intuitive view of action and the findings of neurophysiology if it turned out that even our most reflective decisions are typically physiologically "fixed" at a point in time when we feel ourselves to be consciously contemplating them. But given the implausibility of epiphenomenalism, and the evident utility of conscious reflection in our lives, we consider this scenario to be extremely unlikely (cf. [3, pp.42–3]).

3 Agency from the Third-Personal Perspective

Returning to the motivation that introduced this paper, suppose we are looking around us, trying to make sense of what we see in the world. We see a wide range of processes generating continual change, many of these closely associated with specific objects or systems. What standpoints can we adopt to try to understand these processes? One possibility is to understand the behaviour of a system with reference to what the philosopher Daniel Dennett calls the *physical stance* [12, p.36]. Put simply, the idea of the physical stance is to start with some original configuration, and then use known laws of nature (physics, chemistry etc.) to predict how this system will behave:

> When I predict that a stone released from my hand will fall to the ground, I am using the physical stance. [...] I attribute mass, or weight, to the stone, and rely on the law of gravity to yield my prediction. [12, p.37]

While the physical stance works well for simple cases such as this, it is of course not practicable for understanding or predicting the behaviour of people, who are far too complex to be understood in this way.

Another possibility is the *design stance*, which involves prediction of behaviour based on our understanding of the purpose that a system is supposed to fulfil. Dennett gives the example of an alarm clock [12, pp.37–39]. When someone presents us with an alarm clock, we do not need to make use of physical laws in order to understand its behaviour. If we know it to be a clock, then we can confidently interpret the numbers it displays as the time, because clocks are designed to display

the time. Likewise, if the clock makes a loud and irritating noise, we can interpret this as an alarm that was set at a specific time, because making loud and irritating noises at specified times (but not otherwise) is again something that alarm clocks are designed to do. No understanding of the clock's internal mechanism is required for such an interpretation (at least in normal cases) — it is justified sufficiently by the fact that alarm clocks are designed to exhibit such behaviour.

Importantly, adopting the design stance towards some system does not require us to consider it as *actually* designed, especially in the light of evolutionary theory. Many aspects of biological systems are most easily understood from a design perspective, in terms of the adaptive functions that the various processes perform in the life and reproduction of the relevant organism, treating these processes (at least to a first approximation) *as though* they had been designed accordingly. The same can also apply to adaptive computer systems, whose behaviour is self-modifying through genetic algorithms or other broadly evolutionary methods. Understanding such systems involves the design stance at two distinct levels: at the first level, their overt behaviour — like that of biological systems — may be most easily predicted in terms of the appropriate functions; while at the second level, the fact that they exhibit such functional behaviour is explicable by their having been *designed* to incorporate the relevant evolutionary mechanisms.

A third possible explanatory stance, and the one that most interests us here, is what Dennett calls the *intentional stance* [11]. From this perspective, we attribute *mental states* to entities and then use a common-sense theory of these mental states to predict how the entity will behave, under the assumption that it makes choices in accordance with its attributed beliefs and desires. The most obvious rationale for this approach is that when explaining human activity, it is often useful to make statements such as the following:

> Janine *believes* it is going to rain.
> Peter *wants* to finish his marking.

These statements make use of a *folk psychology*, by which human behaviour is predicted and explained through the attribution of *attitudes*, such as believing and wanting, hoping, fearing, and so on (see, for example, [35] for a discussion of folk psychology). This style of explanation is entirely commonplace, and most people reading the above statements would consider their meaning to be entirely clear, without a second glance.

Notice that the attitudes employed in such folk psychological descriptions are *intentional* notions: they are directed towards some form of *propositional content*. In the above examples, the propositional content is respectively something like "it is going to rain" and "finish my marking", but although it is surprisingly hard to pin down how such content should be characterised or individuated (especially when it involves the identification or possibly misidentification of objects from different perspectives [26, §5]), we need not worry here about the precise details. Dennett coined the term *intentional system* to describe entities

> whose behaviour can be predicted by the method of attributing belief, desires and rational acumen [11, p.49]

The intentional stance can be contrasted not only with the physical and design stances, but also with the *behavioural* view of agency. The behavioural view — most famously associated with B. F. Skinner — attempts to explain human action in terms of stimulus-response behaviours, which are produced via "conditioning" with positive and negative feedback. But as Pinker critically remarks,

> The stimulus-response theory turned out to be wrong. Why did Sally run out of the building? Because she believed it was on fire and did not want to die. [...] What [predicts] Sally's behaviour, and predicts it well, is whether she *believes* herself to be in danger. Sally's beliefs are, of course, related to the stimuli impinging on her, but only in a tortuous, circuitous way, mediated by all the rest of her beliefs about where she is and how the world works. [29, pp.62–63]

In practice, then, the intentional stance is indispensable for our understanding of other humans' behaviour. But it can also be applied, albeit often far less convincingly, to a wide range of other systems, many of which we certainly would not wish to admit as autonomous agents. For example, consider a conventional light switch, as described by Shoham:

> It is perfectly coherent to treat a light switch as a (very cooperative) agent with the capability of transmitting current at will, who invariably transmits current when it believes that we want it transmitted and not otherwise; flicking the switch is simply our way of communicating our desires. [32, p.6]

However, the intentional stance does not seem to be an *appropriate* way of understanding and predicting the behaviour of light switches: here it is far simpler to adopt the physical stance (especially if we are manufacturing light switches) or the design stance (if we are an ordinary user, needing to know only that the switch is designed to turn a light on or off). By contrast, notice that, at least as sketched by Shoham, an intentional explanation of the switch's behaviour requires the attribution to it of quite complex representational states, capable of representing not only the flowing or absence of current, but also our own desires (which, on this story, it acts to satisfy). So even if this intentional account provides accurate prediction of the switch's behaviour, it is wildly extravagant as an *explanation*: to attribute beliefs and desires to a switch is already implausible, but to attribute to it *higher-order* beliefs and desires is well beyond the pale.

4 Higher-Order Intentionality

Human beings are in the unusual position of being both intentional agents in the first-personal sense and also fertile ascribers of third-personal intentionality to other entities. Although above we have described the intentional stance as a third-person explanatory framework, that stance is not of course employed only by people of scientific inclination: indeed the intentional stance comes very naturally — and often far *too* naturally [27, §1] — to people in general.

This human predilection for the intentional stance seems to be intimately bound to our status as *social* animals. That is, the adaptive role of such intentional ascription seems to be to enable us to understand and predict the behaviour of other agents in society. In navigating our way through this complex social web, we become involved in *higher-order* intentional thinking, whereby the plans of individuals (whether ourselves or those we observe) are influenced by the anticipated intentional behaviour of other agents. The value of such thinking is clear from its ubiquity in human life and the extent to which we take it for granted in our communications. Take for example the following fragment of conversations between Alice and Bob (attributed by Baron-Cohen [2] to Pinker):

> *Alice*: I'm leaving you.
> *Bob*: Who is he?

The obvious intentional stance explanation of this scenario is simple, uncontrived, and compelling: Bob *believes* that Alice *prefers* someone else to him and that she is *planning* accordingly; Bob also *wants* to *know* who this is (perhaps in the *hope* of *dissuading* her), and he *believes* that asking Alice will *induce* her to tell him. It seems implausibly difficult to explain the exchange *without* appealing to concepts like belief and desire, not only as playing a role in the agents' behaviour, but also featuring explicitly in their own thinking and planning.

Adoption of the third- (or second-) person intentional stance is also a key ingredient in the way we *coordinate* our activities with each other on a day-by-day basis, as Pinker illustrates:

> I call an old friend on the other coast and we agree to meet in Chicago at the entrance of a bar in a certain hotel on a particular day two months hence at 7:45pm, and everyone who knows us predicts that on that day at that time we will meet up. And we do meet up. [...] The calculus behind this forecasting is intuitive psychology: the knowledge that I *want* to meet my friend and vice versa, and that each of us *believes* the other will be at a certain place at a certain time and *knows* a sequence of rides, hikes, and flights that will take us there. No science of mind or brain is likely to do better. [29, pp.63–64]

All of this involves a mix of first- and higher-order intentional ascription, characterised by Dennett as follows:

> A *first-order* intentional system has beliefs and desires (etc.) but no beliefs and desires *about* beliefs and desires. [...] A *second-order* intentional system is more sophisticated; it has beliefs and desires (and no doubt other intentional states) about beliefs and desires (and other intentional states) — both those of others and its own. [11, p.243]

The following statements illustrate these different levels of intentionality:

1st order: Janine *believed* it was raining.
2nd order: Michael *wanted* Janine to *believe* it was raining.
3rd order: Peter *believed* Michael *wanted* Janine to *believe* it was raining.

In our everyday lives, it seems we probably do not use more than about three layers of the intentional stance hierarchy (unless we are engaged in an artificially constructed intellectual activity, such as solving a puzzle or complex game theory), and it seems that most of us would probably struggle to go beyond fifth order reasoning.

Interestingly, there is some evidence suggesting that other animals are capable of and make use of at least *some* higher-order intentional reasoning. Consider the example of vervet monkeys [10], which in the wild make use of a warning cry indicating to other monkeys the presence of leopards (a threat to the monkey community):

> Seyfarth reports (in conversation) an incident in which one band of vervets was losing ground in a territorial skirmish with another band. One of the losing-side monkeys, temporarily out of the fray, seemed to get a bright idea: it suddenly issued a leopard alarm (in the absence of any leopards), leading *all* the vervets to take up the cry and head for the trees — creating a truce and regaining the ground his side had been losing. [...] If this act is not just a lucky coincidence, then the act is truly devious, for it is not simply a case of the vervet uttering an imperative "get into the trees" in the expectation that *all* the vervets will obey, since the vervet should not expect a rival band to honor *his* imperative. So either the leopard call is [...] a *warning* — and hence the utterer's credibility but not authority is enough to explain the effect, or our utterer is more devious still: he *wants* the rivals to *think* they are *overhearing* a command *intended* only for his own folk. [10, p.347]

One can, of course, put forward alternative explanations for the above scenario, which do not imply any higher-order intentional reasoning. But, nevertheless, this anecdote (amongst others) provides tentative support for the claim that some non-human animals engage in higher-order intentional reasoning. There are other examples: chimpanzees, for example, seem to demonstrate some understanding of how others see them, a behaviour that is indicative of such higher-order reasoning.

4.1 Higher-Order Intentionality and Species Intelligence

While there is evidence that some other animals are capable of higher-order intentional reasoning to a limited extent, there seems to be no evidence that they are capable of anything like the richness of intentional reasoning that humans routinely manage. Indeed, it is tempting to take the widespread ability to reason at higher orders of intentionality as a general indicator of species intelligence. This idea, as we shall see, can be given further support from Robin Dunbar's work on the analysis of social group size in primates [13].

Dunbar was interested in the following question: Why do primates have such large brains (specifically, neocortex size), compared with other animals? Ultimately, the brain is an (energetically expensive) information processing device, and so a large brain would presumably have evolved to deal with some important information processing requirement for the primate. But what requirement, exactly? Dunbar considered a number of primates, and possible factors that might imply the need for enhanced information processing capacity. For example, one possible explanation could be the need to keep track of food sources in the primate's environment. Another possible explanation could be the requirement by primates with a larger ranging or foraging area to keep track of larger spatial maps. However, Dunbar found that the factor that best predicted neocortex size was the primate's *mean group size*: the average number of animals in social groups. This suggests that the large brain

size of primates is needed to keep track of, maintain, and exploit the social relationships in primate groups.

Dunbar's research suggests a tantalising question: given that we know the average human neocortex size, what does his analysis predict as being the average group size for humans? The value obtained by this analysis is now known as *Dunbar's number*, and it is usually quoted as 150. That is, given the average human neocortex size and Dunbar's analysis of other primates, we would expect the average size of human social groups to be around 150. Dunbar's number would remain a curiosity but for the fact that subsequent research found that this number has arisen repeatedly, across the planet, in terms of human social group sizes. For example, it seems that neolithic farming villages typically contained around 150 people. Of more recent interest is the fact that Dunbar's number has something to say about Internet-based social networking sites such as FaceBook. We refer the reader to [14] for an informal discussion of this and other examples of how Dunbar's number manifests itself in human society.

If species neocortex size does indeed correlate strongly with social group size, then the most likely evolutionary explanation seems to be precisely the need for, and adaptive value of, higher-order intentional reasoning within a complex society. Whether hunting in groups, battling with conflicting tribes, pursuing a mate (perhaps against rivals), or gaining allies for influence and leadership (with plentiful potential rewards in evolutionary fitness), the value of being able to understand and anticipate the thinking of other individuals is obvious. We have already seen how higher-order intentional reasoning plays an important role in relationships between humans, to the extent that we routinely take such reasoning for granted in mutual communication. This being so, it is only to be expected that larger social groups would make more demands of such reasoning, providing an attractive explanation for the relationship with neocortex size that Dunbar identified (cf. his discussion in [14, p.30]). This is further corroborated by evidence that higher-order intentional reasoning capabilities are approximately a linear function of the relative size of the frontal lobe of the brain [14, p.181], which seems to be peculiar to primates, and is generally understood as that part of the brain that deals with conscious thought.

5 The Human Intentional System

In this section, we briefly review a model of the *human* intentional system. The model was proposed by Simon Baron-Cohen [2], an evolutionary psychologist interested in understanding of what he calls "mindreading" — the process by which humans understand and predict each other's mental states. A particular interest of Baron-Cohen's is the condition known as autism, which we will discuss in more detail below.

Baron-Cohen's model of the human intentional system is composed of four main modules — see Figure 1. Broadly speaking, the model attempts to define the key mechanisms involved in going from observations of processes and actions in the

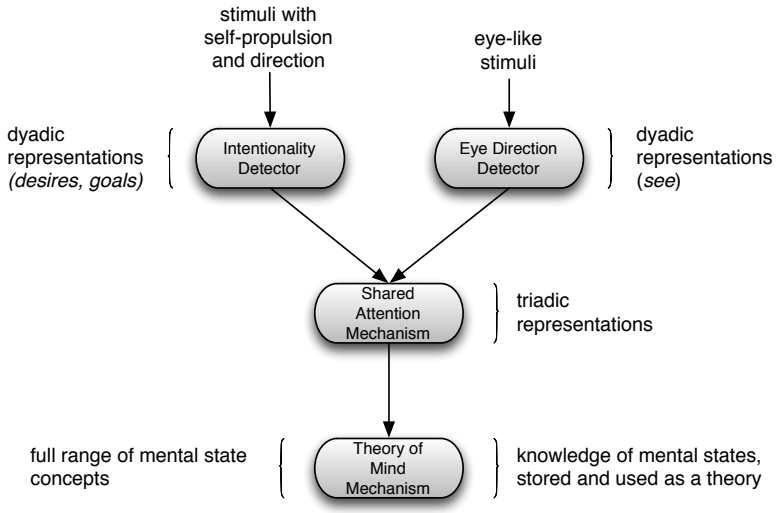

Fig. 1 Baron-Cohen's model of human intentional systems [2, p.32].

environment, through to predictions and explanations of agent behaviour. The four components of the model are as follows:

- the *Intentionality Detector (ID)*;
- the *Eye Direction Detector (EDD)*;
- the *Shared Attention Mechanism (SAM)*; and
- the *Theory of Mind Mechanism (ToMM)*.

The role of the Intentionality Detector (ID) is to:

> [I]nterpet motion stimuli in terms of the primitive volitional mental states of goal and desire. [...] This device is activated whenever there is any perceptual input that might identify something as an agent. [...] This could be anything with self-propelled motion. Thus, a person, a butterfly, a billiard ball, a cat, a cloud, a hand, or a unicorn would do. Of course, when we discover that the object is not an agent — for example, when we discover that its motion is not self-caused, we can revise our initial reading. The claim, however, is that we readily interpret such data in terms of the object's goal and/or desire. [...] ID, then, is very basic. It works through the senses (vision, touch, audition), [...and it will] interpret almost anything with self-propelled motion, or anything that makes a non-random sound, as an agent with goals and desires. [2, pp.32–33].

The output of the ID is primitive dyadic (two-place) intentional ascriptions, such as:

- She wants to stay dry.
- It wants to catch the wildebeest.

At broadly the same level as ID in Baron-Cohen's model is the Eye Direction Detector (EDD). In contrast to ID, which works on multiple types of perceptual input, the EDD is focussed around vision. The basic role is as follows:

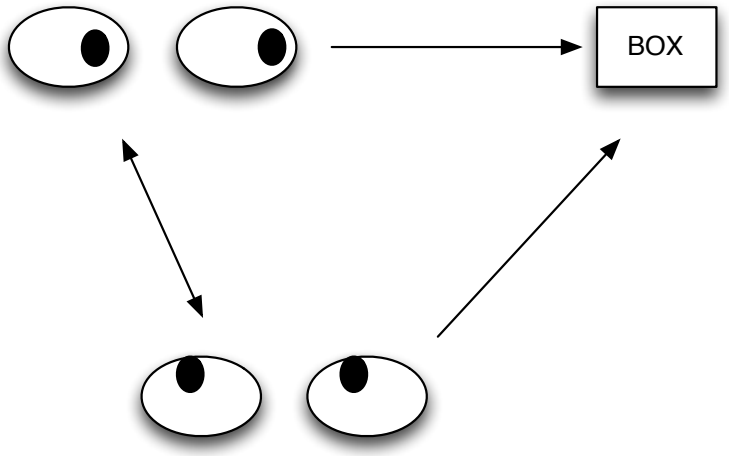

Fig. 2 The SAM builds triadic representations, such as "you and I see that we are looking at the same object" [2, p.45].

> EDD has three basic functions: it detects the presence of eyes or eye-like stimuli, it computes whether eyes are directed towards it or toward something else, and it infers from its own case that if another organism's eyes are directed at something then that organism sees that thing. The last function is important because it [makes it possible to] attribute a perceptual state to another organism (such as "Mummy sees me"). [2, pp.38-39]

Dyadic representations such as those above provide a foundation upon which richer intentional acriptions might be developed, but they simply capture an attitude that an agent has to a proposition, and in this they are of limited value for understanding multi-agent interactions. The purpose of the Shared Attention Mechanism (SAM) is to build *nested*, triadic representations. Figure 2 illustrates a typical triadic representation: "You and I see that we are looking at the same object". Other examples of triadic representations include:

- Bob sees that Alice sees the gun.
- Alice sees that Bob sees the girl.

The Theory of Mind Mechanism (ToMM) is the final component of Baron-Cohen's model:

> ToMM is a system for inferring the full range of mental states from behaviour — that is, for employing a "theory of mind". So far, the other three mechanisms have got us to the point of being able to read behaviour in terms of *volitional mental states* (desire and goal), and to read eye direction in terms of *perceptual mental states* (e.g., see). They have also got us to the point of being able to verify that different people can be experiencing these particular mental states about the same object or event (shared attention). But a theory of mind, of course, includes much more. [2, p.51]

Thus, the ToMM goes from low-level intentional ascriptions to richer nested models. It is the ToMM to which we must appeal in order to understand Bob's question "Who is he?" when Alice says "I'm leaving you."

Baron-Cohen probably does not intend us to interpret the word "theory" in the ToMM to mean a theory in any formal sense (e.g., as a set of axioms within some logical system). However, this suggests an intriguing research agenda: To what extent can we come up with a *logical* ToMM, which can model the same role as the ToMM that we all have? While progress has been made on studying *idealised* aspects of *isolated* components of agency, such as knowledge [21, 15], attempting to construct an *integrated theory of agency* is altogether more challenging. We refer the reader to [9, 38, 37] for discussion and detailed references.

5.1 Autism

It is illuminating to consider what the consequences would be if some of the mechanisms of a fully-fledged intentional system were damaged or malfunctioning. Baron-Cohen hypothesises that the condition known as *autism* is a consequence of impairments in the higher-order mechanisms of the human intentional system: the SAM and/or ToMM. Autism is a serious, widespread psychiatric condition that manifests itself in childhood:

> The key symptoms [of autism] are that social and communication development are clearly abnormal in the first few years of life, and the child's play is characterized by a lack of the usual flexibility, imagination, and pretense. [...] The key features of the social abnormalities in autism [...] include lack of eye contact, lack of normal social awareness or appropriate social behaviour, "aloneness", one-sideness in interaction, and inability to join a social group. [2, pp.62–63]

Baron-Cohen argues that autism is the result of failures in the higher-order components of the human intentional system described above, i.e., those mechanisms that deal with triadic representations and more complex social reasoning: the SAM and ToMM. He presents experimental evidence to support the claim that the ID and EDD mechanisms are typically functioning normally in children with autism [2]. For example, they use explanations such as "she wants an ice cream" and "he is going to go swimming" to explain stories and pictures, suggesting that the ID mechanism is functioning (recall that the role of the ID mechanism is to interpret apparently purposeful actions in terms of goals and desires). Moreover, they are able to interpret pictures of faces and make judgements such as "he is looking at me", suggesting that the EDD mechanism is functioning. However, autistic children seem unable to engage in shared activities, such as pointing to direct the gaze of another individual, suggesting that the SAM is not functioning properly. Finally, experiments indicate that autistic children have difficulty reasoning about the mental states of others, for example, trying to understand what others believe and why. Baron-Cohen takes this as a failure of the ToMM.

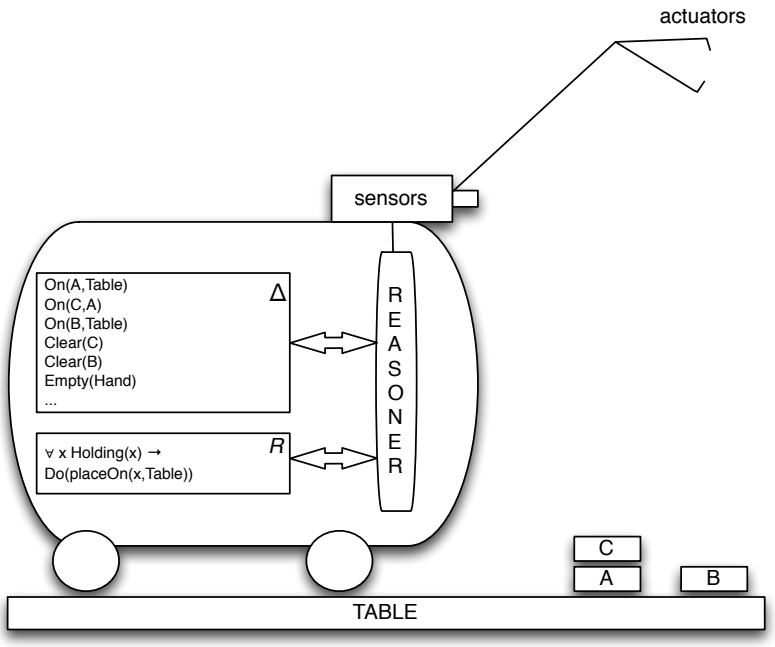

Fig. 3 An artificial agent that decides what to do via logical reasoning.

To evaluate Baron-Cohen's theory, consider how individuals with an impaired higher-order intentional system would behave. We might expect them to have difficulty in complex social settings and in predicting how others will react to their actions, to struggle when attempting to engage in group activities; and so on. And indeed, it seems these behaviours correlate well with the observed behaviours of autistic children.

6 Agency and Artificial Intelligence

Our discussion thus far has been divorced from the question of how we might actually *build* computer systems that can act as autonomous agents, and how far consideration of the nature of human agency can yield insights into how we might go about doing this. This question is of course central to the discipline of artificial intelligence — indeed one plausible way of defining the aim of the artificial intelligence field is to say that it is concerned with building artificial autonomous agents [30].

We start by considering the *logicist* tradition within artificial intelligence, which was historically very influential. It dates from the earliest days of artificial intelligence research, and is perhaps most closely associated with John McCarthy (the

man who named the discipline of artificial intelligence — see, e.g., [24] for an overview of McCarthy's programme). As the name suggests, logic and logical reasoning take centre stage in the logicist tradition, whose guiding theme is that the fundamental problem faced by an agent — that of deciding what action to perform at any given moment — is reducible to a problem of purely logical reasoning. Figure 3 illustrates a possible architecture for a (highly stylized!) logical reasoning agent (cf. [16, pp.307–328]):

- The agent has sensors, the purpose of which is to obtain information about the agent's environment. In contemporary robots, such sensors might be laser range finders, cameras, and radars, and GPS positioning systems [36].
- The agent has effectors, through which it can act upon its environment (e.g., robot arms for manipulating objects, wheels for locomotion).
- The two key data structures within an agent are a set Δ of logical formulae, which represent the *state* of the agent, and a set of rules, R, which represent the *theory* of the agent. The set Δ will typically include information about the agent's environment, and any other information recorded by the agent as it executes. The rule set R will typically include both a *background theory* (e.g., information such as "if an object is on top of a block, then that block is not clear") and a *theory of rational choice* for the agent.
- *Transducers* transform raw sensor data into the symbolic logical form of Δ. Similarly, they map software instructions issued by the robot to commands for the actuators and effectors of the robot.
- A general-purpose *logical reasoning* component enables the agent to apply rules R to the agent's database Δ to derive logical conclusions; we also assume this component handles updating of Δ in the face of new sensor data, etc.

The agent continually executes a *sense-reason-act* loop, as follows:

- *Sense*: The agent observes its environment through its sensors, and after appropriate processing by transducers, this provides potentially new information in logical form; this new information is then incorporated into the agent's representation Δ.
- *Reason*: The reasoning component of the agent then tries to prove a sequent of the form $\Delta \vdash_R Do(\alpha)$, where α is a term that will correspond to an action available to the agent (e.g., an action with the robot arm). The idea is that, if the agent is able to prove such a sequent, then assuming the agent's representation Δ is correct, and the rules R have been constructed appropriately, then α will be the appropriate ("optimal") action for the agent to take.
- *Act*: At this point, the action α selected during the previous phase stage is executed.

Thus, the "program" of the agent is encoded within its rules R. If these rules are designed appropriately, and if the various subsystems of the agent are operating correctly, then the agent will independently select an appropriate action to perform every time it cycles round the sense-reason-act loop.

The idea of building an agent in this way is seductive. The great attraction is that the rules R explicitly encode a theory of rational action for our agent. If the theory

is good, then the decisions our agent makes will also be good. However, there are manifold difficulties with the scheme, chief among them being the following (see, e.g., [4] for a detailed discussion):

- The problem of representing information about complex, dynamic, multi-agent environments in a declarative logical form.
- The problem of translating raw sensor data into the appropriate declarative logical form, in time for this information to be of use in decision-making.
- The problem of automating the reasoning process (i.e., checking whether $\Delta \vdash_R Do(\alpha)$), particularly when decisions are required promptly.

Despite great efforts invested into researching these problems over the past half century, they remain essentially unsolved in general, and the picture we paint above of autonomous decision-making via logical reasoning does not represent a mainstream position in contemporary artificial intelligence research. Indeed, in the late 1980s and early 1990s, many researchers in artificial intelligence began to reject the logicist tradition, and began to look to alternative methods for building agents (see [8] for a detailed discussion of alternative approaches to artificial agency by Rodney Brooks, one of the most prominent and outspoken researchers against the logicist tradition and behind alternative proposals for building agents, and see [38] for a discussion and detailed references).

Before we leave the logicist tradition of artificial intelligence, it is interesting to comment on the status of the logical representation Δ within an agent. The database Δ intuitively contains all the information that the agent has gathered and retained from its environment. For example, referring back to Figure 3, we see that the agent has within its representation Δ the predicate $On(A, Table)$; and we can also see that indeed the block labelled "A" is in fact on top of the table. It is therefore very tempting to interpret Δ as being the *beliefs* of the agent, and thus that *the agent believes block "A" is on the table*. Under this interpretation, the presence of a predicate $P(a,b)$ in Δ would mean that "the agent believes $P(a,b)$", and we would be inclined to say the agent's belief was correct if, when we examined the agent's environment, we found that the object a stood in relation P to object b (this assumes, of course, that we know what objects/relations a, b, and P are supposed to denote in the environment: the agent designer can presumably give us this mapping). See Konolige [22] for a detailed discussion of this subject.

6.1 A Refinement: Practical Reasoning Agents

The practical difficulties in attempting to realise the vision of autonomous agents have led researchers to explore alternatives, and such exploration can also be motivated by the consideration that *we* don't seem to make decisions in that way! While there are surely occasions when many of us use abstract reasoning and problem solving techniques in deciding what to do, it is hard to imagine many realistic situations in which our decision-making is realised via logical proof. An alternative is

to view decision-making in autonomous agents as a process of *practical reasoning*: reasoning directed towards action, rather than beliefs. That is, practical reasoning changes our actions, while theoretical reasoning changes our beliefs [6]:

> Practical reasoning is a matter of weighing conflicting considerations for and against competing options, where the relevant considerations are provided by what the agent desires/values/cares about and what the agent believes. [7, p.17]

Bratman [7] distinguishes two processes that take place in practical reasoning: *deliberation* and *means-ends reasoning*. Deliberation is the process of deciding *what we want to achieve*. As a result of deliberating, we fix upon some *intentions*: commitments to bring about specific states of affairs. Typically, deliberation involves considering multiple possible candidate states of affairs, and choosing between them. The second process in practical reasoning involves determining how to achieve the chosen states of affairs, given the means available to the agent; this process is hence called *means-ends* reasoning. The output of means-ends reasoning is a *plan*: a recipe that can be carried out by the agent, such that after the plan is carried out, the intended end state will be achieved.

Thus, after practical reasoning is completed, the agent will have chosen some intentions, and will have a plan that is appropriate for achieving these intentions. Under normal circumstances, an agent can proceed to execute its chosen plans, and the desired ends will result. The following *practical syllogism* provides a link between beliefs, intentions, plans, and action:

> *If* I intend to achieve ϕ *and*
> I believe plan π will accomplish ϕ
> *Then* I will do π.

The practical reasoning model has been hugely influential within the artificial intelligence community (see, e.g., [1, 18]). A typical architecture for a practical reasoning agent is illustrated in Figure 4. The agent has three key data structures, which, as in the logicist tradition, are symbolic/logical representations. The agent's beliefs are a representation of the agent's environment; the agent's goal represents a state of affairs that the agent is currently committed to bringing about, and the agent's plan is a sequence of actions that the agent is currently executing. If the agent's beliefs are correct, and the plan is sound, then the execution of the plan will result in the accomplishment of the goal [23].

Architectures of the type shown in Figure 4 are often referred to as *belief-desire-intention* (BDI) architectures. In this context, "desire" is usually considered as an intermediate state: the agent has potentially many conflicting desires, but chooses between them to determine the goal or intention that it will then fix on. In the BDI model, the sense-reason-act decision-making loop is modified as follows [37]:

- *Sense*: Observe the environment, and update beliefs on the basis of observations.
- *Option generation*: Given the current beliefs and intentions of the agent, determine what *options* are available, i.e., those states of affairs that the agent *could* usefully commit to bringing about.

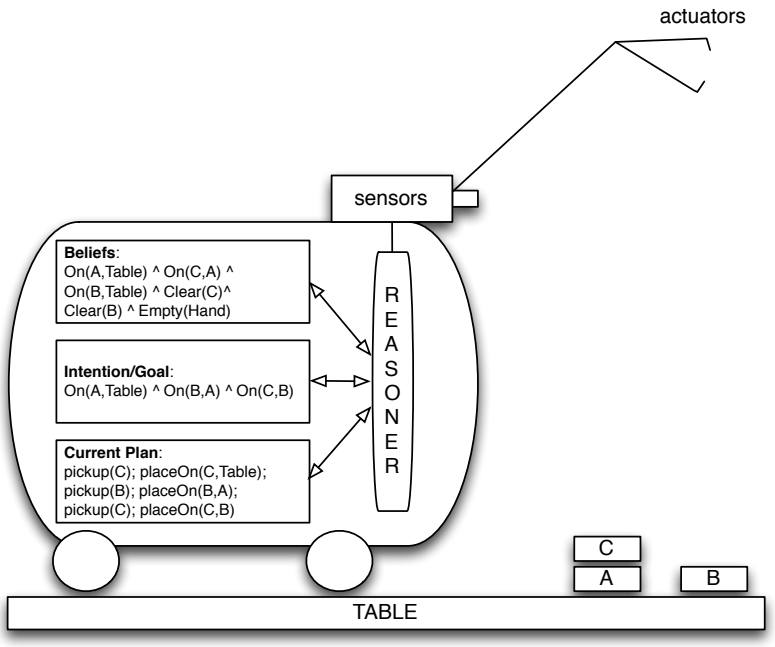

Fig. 4 A practical reasoning agent.

- *Filtering*: Given the current beliefs, desires, and intentions of the agent, choose between competing options and commit to one. The chosen option becomes the agent's current *intention*.
- *Means-Ends Reasoning*: Given the current beliefs and intentions of the agent, find a plan such that, when executed in an environment where the agent's beliefs are correct, the plan will result in the achievement of the agent's intentions.
- *Action*: Execute the plan.

Various refinements can be made to this loop (e.g., so that an agent is not assumed to execute the entire plan before observing its environment again) [37], and of course the picture can be complicated considerably to take account of uncertainties and interaction with complex and changing situations (including game-theoretical consideration of the planning and behaviour of other agents).

The practical reasoning/BDI paradigm suffers from many of the same difficulties that beset the logicist paradigm. For example, the assumption of a logical representation of the agent's beliefs implies the need for transducers that can obtain logical representations from raw sensor data. In addition, the means-ends reasoning problem is computationally complex for logical representations of even modest richness [18]. However, various refinements permit efficient implementations of the architecture; perhaps the best known is the *reactive planning* class of architectures [17, 5]. The basic idea in such architectures is that the agent is equipped with

a collection of plans, generated by the agent designer, which are labelled with the goals that they can be used to achieve. The means-ends reasoning problem then reduces to the comparatively tractable problem of searching through the plan library to try to find a plan that is suitable for the current intention.

Of course, one could now ask to what extent such an agent is genuinely *autonomous*, when in a sense "all it is doing" is assembling and executing plans made up of pre-compiled plan fragments. Such questions raise deep philosophical issues that we cannot address fully now, but here is a sketch of a response. First, there is much to be said for the idea that autonomy is a matter of degree: everything that we do is subject to constraints of various sorts (of ability, cost, law, physical possibility etc.), and all of these can vary in countless ways that extend or limit how far things are "under our control". Secondly, the autonomy attributable to a human — and by extension a computer system — depends in part on how far the reasoning employed in deciding on a course of action is "internal" to the agent: if the agent is performing a complex calculation, taking into account the various constraints and aims within a range of flexible possibilities, this demonstrates far more autonomy than an agent that is simply following orders without any internal reasoning or selection of choices. Thirdly, it follows that autonomy correlates quite strongly with the extent to which application of the third-person intentional stance assists in deep understanding and prediction of the system's behaviour. Some researchers, inspired by the utility of this view of such systems, have proposed the idea of *agent-oriented programming*, in which intentional stance notions such as belief, desire, and intention are first-class entities in a programming language for autonomous agents [33].

7 Conclusions

Our discussion has revealed many ways in which research on agency has led to a convergence in our understanding of human and of artificial agents. In both, the folk-psychological intentional stance — in terms of attributed beliefs and desires — has clear predictive value, and in both, our attempts at a deeper understanding that goes beyond folk psychology have led to plan-based models that shed light on our own behaviour, and also point the way towards practical development of artificial agents. Whether we should describe such an artificial system as a genuine "agent" and as having *literal* "beliefs" and "desires" is, of course, a matter for debate. But when a system is sufficiently sophisticated that its behaviour can only feasibly be understood or predicted in terms of belief-like, desire-like, and plan-like states and the interplay between those (rather than purely in terms of simple execution of prepackaged instructions), we think there is a lot to be said for extending the boundaries of our concepts accordingly. As Millican [27, §3–4] has argued in the case of artificial intelligence, the development of such systems has faced us with a new problem which is not anticipated by the established boundaries of our traditional concepts. The concept of "agency", like that of "intelligence" is *open textured*, and how we

mould it within this new context is largely a matter of decision, rather than mere analysis of our pre-existing conceptual repertoire.

There are several possible avenues for future research. One interesting open problem remains the extent to which we can develop *formal* theories that can predict and explain the behaviour of *human* agents; another is the extent to which we can link such formal theories with computer programs, in order to provide an account of their behaviour in terms of agentive concepts such as beliefs, desires, and rational choice.

Acknowledgements

Wooldridge was supported by the European Research Council under Advanced Grant 291528 ("RACE").

References

1. J. F. Allen, J. Hendler, and A. Tate, editors. *Readings in Planning*. Morgan Kaufmann Publishers: San Mateo, CA, 1990.
2. S. Baron-Cohen. *Mindblindness: An Essay on Autism and Theory of Mind*. The MIT Press: Cambridge, MA, 1995.
3. T. Bayne. Libet and the case for free will scepticism. In R. Swinburne, editor, *Free Will and Modern Science*, pages 25–46. British Academy, 2011.
4. L. Birnbaum. Rigor mortis. In D. Kirsh, editor, *Foundations of Artificial Intelligence*, pages 57–78. The MIT Press: Cambridge, MA, 1992.
5. R. Bordini, J. F. Hübner, and M. Wooldridge. *Programming multi-agent systems in AgentSpeak using Jason*. John Wiley & Sons, 2007.
6. M. E. Bratman. *Intention, Plans, and Practical Reason*. Harvard University Press: Cambridge, MA, 1987.
7. M. E. Bratman. What is intention? In P. R. Cohen, J. L. Morgan, and M. E. Pollack, editors, *Intentions in Communication*, pages 15–32. The MIT Press: Cambridge, MA, 1990.
8. R. A. Brooks. *Cambrian Intelligence*. The MIT Press: Cambridge, MA, 1999.
9. P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.
10. D. C. Dennett. Intentional systems in cognitive ethology. *The Behavioral and Brain Sciences*, (6):343–390, 1983.
11. D. C. Dennett. *The Intentional Stance*. The MIT Press: Cambridge, MA, 1987.
12. D. C. Dennett. *Kinds of Minds*. London: Phoenix, 1996.
13. R. I. M. Dunbar. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22:469–493, 1992.
14. R. I. M. Dunbar. *How Many Friends Does One Person Need?: Dunbar's Number and Other Evolutionary Quirks*. Faber and Faber, 2011.
15. R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning About Knowledge*. The MIT Press: Cambridge, MA, 1995.
16. M. R. Genesereth and N. Nilsson. *Logical Foundations of Artificial Intelligence*. Morgan Kaufmann Publishers: San Mateo, CA, 1987.
17. M. P. Georgeff and A. L. Lansky. Reactive reasoning and planning. In *Proceedings of the Sixth National Conference on Artificial Intelligence (AAAI-87)*, pages 677–682, Seattle, WA, 1987.

18. M. Ghallab, D. Nau, and P. Traverso. *Automated Planning: Theory and Practice*. Morgan Kaufmann Publishers: San Mateo, CA, 2004.
19. A. I. Goldman. Action. In S. Guttenplan, editor, *A Companion to the Philosophy of Mind*, pages 117–121. Blackwell, 1995.
20. S. Guttenplan, editor. *A Companion to the Philosophy of Mind*. Blackwell, 1995.
21. J. Hintikka. *Knowledge and Belief*. Cornell University Press: Ithaca, NY, 1962.
22. K. Konolige. *A Deduction Model of Belief*. Pitman Publishing: London and Morgan Kaufmann: San Mateo, CA, 1986.
23. V. Lifschitz. On the semantics of STRIPS. In M. P. Georgeff and A. L. Lansky, editors, *Reasoning About Actions & Plans — Proceedings of the 1986 Workshop*, pages 1–10. Morgan Kaufmann Publishers: San Mateo, CA, 1986.
24. J. McCarthy. *Formalization of common sense: papers by John McCarthy*. Ablex Publishing Corp., 1990.
25. A. R. Mele and Paul K. Moser. Intentional action. *Nous*, 28(1):39–68, 1994.
26. P. Millican. Content, thoughts, and definite descriptions. *Proceedings of the Aristotelian Society, Supplementary Volume*, 64:167–203, 1990.
27. P. Millican. The philosophical significance of the Turing machine and the Turing test. In S. B. Cooper and J. van Leeuwen, editors, *Alan Turing: His Work and Impact*, pages 587–601. Elsevier, 2013.
28. D. A. Norman and T. Shallice. Attention to action; willed and automatic control of behaviour. In R. J. Davidon, G. E. Schwartz, and D. Shapiro, editors, *Consciousness and Self-Regulation: Advances in Research and Theory Volume 4*, pages 1–18. Plenum Press, 1986.
29. S. Pinker. *How the Mind Works*. W. W. Norton & Co., Inc.: New York, 1997.
30. S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, 1995.
31. J. R. Searle. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press: Cambridge, England, 1983.
32. Y. Shoham. Agent-oriented programming. Technical Report STAN–CS–1335–90, Computer Science Department, Stanford University, Stanford, CA 94305, 1990.
33. Y. Shoham. Agent-oriented programming. *Artificial Intelligence*, 60(1):51–92, 1993.
34. C. S. Soon, M. Brass, H.-J. Heinze, and J.-D. Haynes. Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, 11(5):543–545, 2008.
35. S. P. Stich. *From Folk Psychology to Cognitive Science*. The MIT Press: Cambridge, MA, 1983.
36. S. Thrun, M. Montemerlo, H. Dahlkamp, D. Stavens, A. Aron, J. Diebel, P. Fong, J. Gale, M. Halpenny, G. Hoffmann, K. Lau, C. Oakley, M. Palatucci, V. Pratt, P. Stang, S. Strohband, C. Dupont, L.-E. Jendrossek, C. Koelen, C. Markey, C. Rummel, J. van Niekerk, E. Jensen, P. Alessandrini, G. Bradski, B. Davies, S. Ettinger, A. Kaehler, A. Nefian, and P. Mahoney. Stanley: The robot that won the DARPA grand challenge. In M. Buehler, K. Iagnemma, and S. Singh, editors, *The 2005 DARPA Grand Challenge*, pages 1–43. Springer-Verlag: Berlin, Germany, 2007.
37. M. Wooldridge. *Reasoning about Rational Agents*. The MIT Press: Cambridge, MA, 2000.
38. M. Wooldridge and N. R. Jennings. Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, 10(2):115–152, 1995.