discriminate legal from illegal bigrams. People are good at this discrimination task, which has been taken to indicate that their knowledge is conscious. The problem is that both conscious and unconscious knowledge would in principle allow such discrimination.

Another method for determining the conscious status of knowledge states is to ask participants to report or discriminate not states of the world (e.g. 'can this bigram occur in the grammar?') but the relevant mental states involved in classification (see OBJECTIVE VS SUBJECTIVE MEASURES OF CONSCIOUSNESS). Unconscious knowledge is knowledge a person is not aware of. Thus, we need to determine whether or not the person knows that they know in order to determine if the knowing is conscious. For example, confidence ratings can be elicited after each classification decision. According to the *guessing criterion*, unconscious knowledge is shown when the participant says they are purely guessing but they are in fact performing above baseline. According to the *zero correlation criterion*, knowledge is unconscious if the person cannot discriminate between when they are guessing and when they have knowledge, i.e. there is no relationship between confidence and accuracy. Both criteria have indicated unconscious knowledge, though a typical pattern is for there to be evidence of both conscious and unconscious knowledge.

The guessing criterion has been criticized because of the bias problem: when people *say* they are guessing, they might *think* that they are not. (Note that the existence of an adjustable bias for *thinking* one is guessing vs knowing is not in itself a problem for the guessing criterion.) A response to the objection has been to indicate evidence that people's reports of whether or not they are guessing distinguish knowledge types that differ in ways predicted by a theory of consciousness (e.g. resilience to a secondary task). The zero correlation criterion is less susceptible to the bias problem.

The guessing and zero correlation criteria measure the conscious status of *judgement knowledge*: i.e. knowing that this string is grammatical. That leaves open the question of whether the person's *structural knowledge* (knowledge of the structure of the training strings) is conscious or unconscious. To address the latter, the experimenter can ask people after each classification decision whether they based their answer on random responding or intuition (unconscious structural knowledge) or rules or memory (conscious structural knowledge). This subjective method indicates that people typically use both conscious and unconscious structural knowledge.

People trained on two grammars in different contexts can choose which of the grammars to use in the classification phase (when the content of their structural knowledge is sufficient for discrimination between the grammars; see MEMORY, PROCESS-DISSOCIATION PROCEDURE). That is, like bilinguals, people can choose which language to use; in that sense, grammatical knowledge is not applied automatically. Further, people trained on one grammar in one context do not apply it in a test phase in a different context unless told of the connection between the contexts. Such intentional control of the use of the knowledge often coexists with lack of awareness of what the knowledge is by the guessing criterion. That is, a person does not need to be aware of controlling their knowledge, nor of what the knowledge is, in order to control it.

A common argument for there being unconscious knowledge learned in artificial grammar learning is that other primates and human babies as young as two months can learn statistical structures in sequences. The assumption made in this argument is that such creatures do not consciously test hypotheses nor do they have fully fledged episodic memory. Further, people with *amnesia caused by damage to the temporal lobes learn artificial grammars almost as well as normal adults. But none of these facts entail that the corresponding learning mechanism in normal human adults produces unconscious knowledge. Moreover, the mere fact that a person has impaired episodic memory does not entail they do not use conscious knowledge, either judgement or structural (e.g. rules). However, these studies on different populations can be very informative about the basis of implicit learning in adults, when its implicit nature is established by other means.

ZOLTÁN DIENES

Dienes, Z. (2008). 'Subjective measures of unconscious knowledge'. In Banerjee, R. and Chakrabarti, B. (eds) *Models of Brain and Mind: Physical, Computational and Psychological Approaches*.

Gómez, R. L. (2006). 'Dynamically guided learning'. In Munakata, Y. and Johnson, M. (eds) *Attention and Performance XXI: Processes of Change in Brain and Cognitive Development*.

Pothos, E. M. (2007). 'Theories of artificial grammar learning'. *Psychological Bulletin*, 133.

Reber, A. S. (1993). *Implicit Learning and Tacit Knowledge: an Essay on the Cognitive Unconscious*.

Shanks, D. R. (2005). 'Implicit learning'. In Lamberts, K. and Goldstone, R. (eds) *Handbook of Cognition*.

artificial intelligence and the study of consciousness Artificial intelligence (AI), in its broadest sense, is any attempt to design and create artefacts that have mental properties, or exhibit characteristic aspects of systems that have such properties. Despite the name, such properties include not just intelligence, but also those having to do with e.g. perception, action, emotion, creativity, and consciousness.

1. Varieties of AI

Although the last half-century or so has seen various approaches to AI, including connectionism/neural networks, dynamical systems engineering, embodied/situated robotics, and artificial life, the term is often used more narrowly to refer to approaches that emphasize symbolic computation. Indeed, it was this particular approach that was dominant among those who first used the term to describe their work (John McCarthy coined the term in 1956), a situation that arguably continues to this day. To avoid confusion in what follows, the term *symbolic artificial intelligence* (or *symbolic AI*) will be used to refer to this specific approach, and *artificial intelligence* (or *AI*) to the general endeavour.

Another distinction can be made between two related, but distinct, goals in pursuing AI. *Engineering AI* is primarily concerned with creating artefacts that can do things that previously only naturally intelligent agents could do; whether or not such artificial systems perform those functions in the way that natural systems do is not considered a matter of primary importance. *Scientific AI*, however, is primarily concerned with understanding the processes underlying mentality, and the technologies provided by engineering AI, however impressive, are only considered of theoretical relevance to the extent that they resemble or otherwise illuminate the mental processes of interest.

Within scientific AI, further distinctions can be made concerning the relation that is believed to hold between the technology involved in an AI system and the mental phenomena being explained. Adapting terminology from Searle (1980), *weak AI* is any approach that makes little or no claim of a relation between the technology and modelled mentality. This would be a use of AI technology in a way similar to the use of computational simulations of hurricanes in meteorology: understanding can be facilitated, but no one supposes that this is because hurricanes are themselves computational in any substantive sense. At the other extreme, *strong AI* is any approach that claims that instantiations of the technologies involved are thereby instantiations of the mental phenomena being explained/modelled. For example, strong symbolic AI maintains that an appropriately programmed computer actually understands, believes, knows, is aware, etc. Between these two poles is what might be termed *moderate AI*. This view, unlike weak AI, claims that the modelling relation holds as a result of deeper, explanatory properties being shared by the AI technology and the mental phenomena being explained/modelled. However, unlike strong AI, moderate AI does not go on to claim that instantiating these common properties is alone sufficient for instantiating those mental phenomena—something else might be required (e.g. in the case of symbolic AI, proper historical/environmental situatedness, or 'symbol grounding'; implementation in living matter as opposed to dead metal and silicon, etc.).

Since a consideration of all the combinations of these approaches is not possible here, the focus will be on the prospects for symbolic, strong, scientific AI with regard to mentality in general. Some perceived limitations of the symbolic approach and proposed alternatives will be discussed before considering the specific issues that arise concerning AI and the understanding of consciousness in particular.

Attempting to use AI to instantiate or explain consciousness is sometimes called **machine consciousness*; see Holland (2003) and the entry for that topic for further discussion, and for some specific examples of work in this area. Other notable examples of AI models of consciousness include the symbolic of Johnson-Laird (1983); the connectionsist of Churchland (1995), Lloyd (1995), and Sun (1999); and the embodied/situated robotics of Dennett (1994).

2. Symbolic AI

The emphasis of symbolic AI is on the processing of representations, specifically symbols. At the heart of the symbolic approach are two features: (1) a sharp distinction between semantic and non-semantic (syntactic) properties; and (2) context-invariant atoms that can be composed, usually concatenatively like language, into complex structures whose syntax and semantics depend systematically on those atoms and their mode of composition.

The programming of digital computers, although thoroughly symbolic, is not in itself a reliable indicator of the symbolic approach to AI, since other approaches often use such technology merely as a means of creating or modelling systems that are (or are claimed to be) non-symbolic or even non-computational.

There have been several motivations for the symbolic approach. One derives from some results in computability theory that pre-date AI and computer technology itself. Turing (1950) introduced a set of formal models of digital, algorithmic computation called *Turing machines, automata that are given a symbol string (usually interpreted as an integer) as an input, and produce a symbol string (also usually interpreted as an integer) as the output for that input. In this way such machines could be understood to be computing functions over the integers. Turing proved that there exist universal Turing

machines that can simulate the action of any other Turing machine. Such machines can therefore be seen to be capable of computing any Turing-computable function. This, coupled with the assumption that if any function can be computed at all, it can be computed by a Turing machine (the *Church–Turing thesis*), yields the result that a universal machine can be seen to be capable of computing anything that is computable at all. In particular, many take this to establish that a universal machine can compute any function a human can compute. If one then adds the assumption that human behaviour, or at least the mental processes that give rise to it, can be conceptualized as mathematical functions of the relevant sort, then it follows that any universal machine can, in principle, be programmed to exhibit behaviour functionally equivalent to that of any human being.

In some sense, this is enough of an existence proof for the purposes of engineering AI, but motivating this approach for scientific AI requires a behaviourist assumption to establish that any such simulation of human behaviour would have, or at least would model, mentality. Since, the Turing test notwithstanding, symbolic AI has anti-behaviourist roots, a different motivation is usually given for scientific applications of symbolic AI: *cognitivism* (or *computationalism*). This is the claim that cognition (more narrowly: thinking; more broadly: all mentality) actually is a kind of (symbolic) computation. It follows from this claim that implementing certain kinds of computation is sufficient for reproducing or modelling cognitive phenomena. Cognitivism is the idea that mentality, in particular thinking, is at root a formal activity: unlike, say, a rainstorm, if you computationally simulate thinking, you actually recreate it. Furthermore, since the steps involved in thinking are, it is assumed, accessible to the thinker, it is in general possible for a thinker to write those steps down, turn them into a set of instructions for a computer, and thereby create a system that reproduces, or at least models, any particular instance of thinking.

3. AI and understanding consciousness

Modelling, in particular computational modelling, confers several benefits on scientific investigation (Sloman 1978), including allowing one to: (1) state theories precisely; (2) test theories rigorously; (3) encounter unforeseen consequences of one's theories; (4) construct detailed causal explanations of actual events; and (5) undergo conceptual change through direct, interactive experience with the phenomena under investigation.

Thus, even at its weakest, AI offers these benefits to the understanding of mentality. Moving from mere simulation toward strong AI (cf. section 1) presumably multiplies these benefits, especially (2) and (3) (cf. Brooks 1991).

There are three ways the AI methodology is usually applied to explaining consciousness. One can attempt to model the physical system underlying consciousness, at a particular level of abstraction (e.g. a connectionist model is pitched at a lower, more hardware-dependent level of abstraction than is a typical symbolic model). One can attempt to model conscious processes directly, by using introspection to note their causal structure, and them implementing this structure in a (usually symbolic) AI system. Or one can attempt to model the behaviour of a system known or believed to be conscious, without any direct knowledge of the underlying physical or phenomenological structure, in the hope that reproducing both actual and counterfactual behaviour is sufficient to ensure that the same consciousness-producing causal structure is thereby implemented.

Central to applying AI methodology to understanding mentality is belief in the multiple realizability of mental states, itself motivated by, e.g., thought experiments concerning creatures ('Martians') that behave just like humans, but have a very different physiology. Since it would be politically incorrect in the extreme to deny mental states to these Martians, it must be a mistake to think that mental states can only be implemented in biological states like those of humans and animals on Earth. The question then arises: what do Martians and Earthlings have in common by virtue of which they both enjoy mental lives? Again, since AI arose out of an anti-behaviourist tradition, the commonality is not believed to be behaviour, but abstract causal organization, something that is describable using computational formalisms such as Turing machines. The belief that it is abstract causal organization that identifies mental states is called *functionalism*; if functionalism is true, then not only is it possible to investigate mentality with non-neural hardware, but also (some would say) it is a mistake to spend much time investigating neurophysiology to explain mentality. Doing so would be like trying, for example, to understand flight by looking at birds' feathers under a microscope. Instead, the analogy continues, we only came to understand natural flight by achieving artificial flight, and we only succeeded in doing that once we stopped trying to copy slavishly the superficial characteristics of biological flyers. Similarly, AI allows one to specify and test the *virtual machine* that, it is proposed, provides the proper level of analysis for explaining mentality.

4. Difficulties

There are several reasons why one might think that AI cannot contribute to the understanding of consciousness.

First, there are the problems shared by all naturalistic approaches. For example, it seems that phenomenal states can be observed directly only by the subject of those states, yet objective or at least inter-subjective observation and verification is thought to be at the very heart of scientific method. There is also the *explanatory gap (Levine 1983), or *hard problem (Chalmers 1996): it seems that naturalistic, non-phenomenal properties of a system do not explain, or at least do not imply, the phenomenal properties of that system. The applicability of these problems to AI explanations of consciousness is most clear in the case of strong AI (cf. section 1): how could one ever know if one has succeeded in creating an artificial consciousness, since one cannot directly observe its purported conscious states? For any AI system that is supposedly conscious, one can always imagine it being built and behaving the same way and yet not being conscious, so what exactly has been explained? Some have argued that these are not the insurmountable problems they seem to be, but there is little consensus on the matter. One constraint can be noted, however: in arguing that we would not be able to know that an AI system is conscious, we should be careful not to set the epistemological bar so high that we call into question our knowledge that other humans are conscious.

Next, there are doubts concerning the ability of AI systems to model cognition/mentality in general. These vary depending on the AI approach; the development of some AI approaches can be seen as attempts to overcome the general limitations of another (usually the symbolic) approach. For example, it is argued that symbolic AI cannot provide an accurate model of human cognition, since in such AI systems, millions of serially dependent operations must be performed to, say, recognize an object, but this is done in the brain in fewer than a hundred such steps. Connectionist AI is then offered as an approach that does not suffer from this problem. Another example of a purported obstacle to symbolic AI in general is the *frame problem* (cf. e.g. Pylyshyn 1987).

Third, there are specific doubts concerning the inability of AI (in particular) to explain consciousness (in particular). These are often aimed specifically at symbolic AI, but there is at least one argument against an AI account of consciousness that applies independently of approach. This argument finds an incompatibility between the possibility of being conscious and at the same time being an artefact in any interesting sense. To be considered artificial, it would seem an AI system would have to be not just the results of our labour (children are that), but also designed by us. But this means that any purpose, meaning, or intentionality in the system is not its own, but rather derivative from our design of it. Yet consciousness, it seems, is autonomous, exhibiting original, non-derivative intentionality. As with all the objections to AI presented here, this argument can be resisted. A sign that it might be too strong is that it would imply that I might not be conscious, since it might be that I was created by design (divine or mundane). Yet surely this is not a possibility I can countenance! It would be odd indeed if the details of my origins, usually believed to be an empirical matter, could be known by me a priori.

Perhaps the most well known of the specific objections to symbolic AI accounts of consciousness is Searle's *Chinese room argument (Searle 1980). Searle argues against the claim that a computer could understand solely by virtue of running the right program. To do this, he exploits the subjective, conscious nature of understanding, and the formal, implementation-independent nature of symbolic computation and programs. Since he can himself implement any purported understanding-endowing program, and presumably would not come to understand anything thereby, he refutes the strong AI claim he targets, or at least appears to do so. It is not necessary here to rehearse the various replies and counter-replies that have been given. But it is of note that what was referred to above as 'moderate AI' (cf. section 1) is immune to this argument. The Chinese room may show that computation is not *sufficient* for conscious understanding, but it does not show that it is not *necessary* for it, nor does it show that computation cannot play an explanatory role with respect to consciousness. And of course the argument does not apply in general to alternative approaches to AI that do not place as much emphasis on implementing formal programs.

Another famous objection to symbolic AI is the *diagonal argument*. This dates back to Gödel and Turing, but was developed philosophically by Lucas (see Lucas 1996 for a retrospect) and, more recently, Penrose (1994). It can be shown that no Turing machine can compute the non-halting function. Enumerate all the Turing machines. Now consider this function: 'For all n, halt if and only if the nth Turing machine does not halt when given input n'. No Turing machine that is sound with respect to this function can halt when given its own number k in the enumeration as an argument (it must halt on k if and only if it does not halt on k). Furthermore, I just proved to you that any such sound Turing machine k does not halt when given input k. Thus you and I can answer this question (compute this function?) correctly for all n, while no Turing machine can. So we can do more than Turing machines. The explanation Penrose offers for this fact is that we are conscious, and can use our consciousness to jump out of the algorithmic 'system', see patterns that are not classically computable, etc. Thus, symbolic AI cannot even match the

performance of a conscious system, let alone explain it or re-instantiate it. As with the Chinese room, there are too many replies and counter-replies to consider them here. But some similar observations can be made; specifically, 'moderate AI' (cf. section 1) is again immune to this argument. Even if there are aspects of consciousness that are non-computational, this does not show that computation of some sort is not necessary/explanatory for those aspects of consciousness. Nor does it show that all aspects or instances of consciousness have non-algorithmic components. Like the Chinese room, the argument does not apply in general to alternative approaches to AI that do not place algorithms centre stage. And one may wonder: even if computers cannot recreate human consciousness, is halting-problem-defying human consciousness the only kind of consciousness possible in this universe? If not, then AI (even symbolic AI) explanations of these other kinds of consciousness have not been shown to be impossible.

5. Unexplored possibilities

It could very well be that there are more possibilities for using AI to understand consciousness than we have yet envisioned. For example, an aspect of AI that had more prominence in the field's early years than it does now is AI as prosthesis: 'artificial intelligence' as a parallel construction to 'artificial leg' rather than 'artificial light'.

Ross Ashby, a venerated pioneer in the field of dynamical AI, proposed a 'design for an intelligence amplifier' (Ashby 1956). Perhaps AI could contribute to our understanding of consciousness as much by systematically altering or extending it as by replicating it. Technologies based on AI may also be required to help us mine and process the enormous quantities of data we anticipate to acquire concerning the operation of the brain over the next decades. If so, AI will have a perhaps more prosaic, though no less crucial, role to play in understanding consciousness.

RON CHRISLEY

Ashby, R. (1956). 'Design for an intelligence-amplifier'. In Shannon, C. E. and McCarthy, J. (eds) *Automata Studies*.

Bechtel, W. (1993). 'Consciousness: perspectives from symbolic and connectionist AI'. *Neuropsychologia*, 33.

Brooks, R. (1991). 'Intelligence without reason'. MIT AI Lab Memo 1293.

Chrisley, R. (ed.) (2000). *Artificial Intelligence: Critical Concepts*. (Many of the earlier references are reprinted here.)

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*.

Churchland, P. M. (1995). *The Engine of Reason, the Seat of the Soul: a Philosophical Journey into the Brain*.

Dennett, D. C. (1994). 'The practical requirements for making a conscious robot'. *Philosophical Transactions of the Royal Society Series A: Physical Sciences*, 349.

Pylyshyn, Z. (ed.) (1987). *The Robot's Dilemma: the Frame Problem in Artificial Intelligence*.

Holland, O. (ed.) (2003). *Machine Consciousness*.

Johnson-Laird, P. N. (1983). *Mental Models*.

Levine, J. (1983). 'Materialism and qualia: the explanatory gap'. *Pacific Philosophical Quarterly*, 64.

Lloyd, D. (1995). 'Consciousness: a connectionist manifesto'. *Minds and Machines*, 5.

Lucas, J. (1996). 'Minds, Machines and Gödel: a Retrospect'. In Millican, P. J. R. and Clark, A. (eds) *Machines and Thought: the Legacy of Alan Turing*.

Penrose, R. (1994). *The Shadows of the Mind: a Search for the Missing Science of Consciousness*.

Searle, J. (1980). 'Minds, brains and programs'. *Behavioral and Brain Sciences*, 3.

Sloman, A. (1978). *The Computer Revolution in Philosophy*.

Sun, R. (1999). 'Accounting for the computational basis of consciousness: a connectionist approach'. *Consciousness and Cognition*, 8.

associative agnosia See AGNOSIA

attentional blink The attentional blink (AB) is a temporary state of poor awareness of current stimuli, lasting about half a second, that is induced by focusing attention and becoming consciously aware of a relevant stimulus object that has just previously been briefly presented.

The concept of the AB was originally developed to describe a phenomenon uncovered in a series of laboratory experiments in which normal human adults were required to report the presence of two different target letters (T1 and T2) that had been embedded within a series of other briefly presented letters, each seen for about 100 ms (Raymond et al. 1992; see Fig. A15a). Although the probability of reporting both targets was found to be very high if the interval (or lag) between their presentations was greater than about 500 ms, the ability to report the second target (T2) dropped precipitously when intervals between 200 and 500 ms were used. In these, and many other studies, it was found that if the T2 item appeared immediately after the T1 item (with no intervening item), no deficit in reporting T2 occurred. This effect has been called *lag-1 sparing*. Figure A15b shows the now classic U-shaped function that relates performance on the T2 task to the T1–T2 interval. The T1–T2 interval has no effect on T1 performance. Critically, if T1 is simply ignored even though it is still presented, no AB (dip in T2 performance) is seen. This is why the effect was called the *attentional blink*: 'attentional' because it is induced by attention to a prior target and 'blink' because it is a normal and temporary period of apparent insensitivity.

Of theoretical interest is the observation that the effect depends on the presentation of a second (or