23 Robots and Machine Consciousness

Antonio Chella

23.1 Introduction

Building a conscious robot is an enormous scientific and technological challenge. Debates about the possibility of sentient robots and the positive outcomes and risks for human beings are no longer confined to philosophical circles. Consciousness is part of the physical world, and therefore its aspects can be studied and even replicated by robot systems.

There is no accepted definition of consciousness so far. Searle (2000) claimed that "consciousness consists of inner, qualitative, subjective states and processes of sentience or awareness. Consciousness, so defined, begins when we wake in the morning from a dreamless sleep and continues until we fall asleep again, die, go into a coma, or otherwise become 'unconscious'" (559). Vimal (2009) overviewed several meanings of the word employed in scientific works related to the study of consciousness.

Although there are contrasting philosophical positions concerning consciousness (see, e.g., Blackmore and Troscianko [2018] for an up-to-date review), it is useful to point out the broad distinction of consciousness as *experience* versus consciousness as *function*. For experience, a subject is conscious when they feel visual experiences, bodily sensations, mental images, and emotions (Chalmers 1995). As Nagel (1974) pointed out, a subject has a conscious experience if there is something that is like to be that subject.

For function, a conscious subject can integrate information (Tononi 2008); they process information that is globally available (Dehaene et al. 2017); they are introspectively aware of themselves (Floridi 2005). Moreover, they possess an inner model of themselves and of the external environment (Holland 2003b). They can anticipate perceptual and behavioral activities (Hesslow 2002). They generate inner speech (Morin 2005) and act by sensorimotor interactions with the external world (O'Regan and Noë 2001), among other capabilities.

In brief, the multidisciplinary effort of robot and machine consciousness is aimed at investigating consciousness in the light of robotics and artificial systems, psychology, philosophy of mind, ethics, and neuroscience. The broad scopes of robot and machine consciousness are:

- to build robots that show forms of functional consciousness by taking inspiration from biological consciousness;
- to build robots based on theoretical issues of consciousness;

- to employ robots as tools to model and to understand biological aspects of consciousness;
- to study procedures aimed at measuring consciousness in robots;
- to discuss ethical problems emerging through the overlap of robotics and consciousness.

23.2 A Brief History of Robot Consciousness

To the best of the author's knowledge, the first occurrence of the word "artificial consciousness" is found in the book *Cybernetic Machines* by T. N. Nemes, published in Hungary in 1962. The book was translated into English in 1970. Nemes, in this early attempt, considered artificial consciousness as the capability of a robot to discriminate between self and others. The author proposed a conceptual sketch of a circuit able to distinguish between proprioceptive inputs that generate sentences as "I go" from shape recognition and motion perception circuits that process data from external inputs able to create sentences as "Peter goes."

The modern scientific framework of artificial and robot consciousness has been primarily introduced by Igor Aleksander (1992, 2015). At the ICANN 1992 Conference in Brighton, Aleksander presented a paper on capturing consciousness in neural systems, where he proposed the postulates defining a conscious organism that may be applied to a biological organism or an artifact. Notably, during the invited talk, Aleksander announced that the "hunting season of artificial consciousness is open."

Another influential early model for machine consciousness is due to Schmidhuber (1992). He discussed machine consciousness by presenting an unsupervised neural network able to discover and learn unexpected events.

The symposium on "Can a Machine Be Conscious," organized by the Swartz Foundation in 2001, was another milestone for robot consciousness. The concluding remarks of Christof Koch, valid still today, stated that "we know of no fundamental law or principle operating in this universe that forbids the existence of subjective feelings in artifacts designed or evolved by humans."[1]

Since 2001, many conferences, workshops, and special issues of journals have been devoted to the field of robot consciousness. Early works are described in the collections edited by Holland (2003a), Clowes et al. (2007), and Chella and Manzotti (2007b). In 2007, the Association for the Advancement of Artificial Intelligence (AAAI) organized a fall symposium on "AI and Consciousness," with the proceedings edited by Chella and Manzotti (2007a).

Reggia (2013) provided quite an up-to-date review of the field. A collection of recent research papers concerning consciousness in humanoid robots was edited by Chella et al. (2019).

During the summer of 2017, SRI International organized a series of workshops on technology and consciousness. The workshops provided a general view of machine consciousness; the outcomes are summarized in a technical report edited by Rushby and Sanchez (2018).

A continuous source of information is the *Journal of Artificial Intelligence and Consciousness* (*JAIC*), formerly known as the *International Journal of Machine Consciousness* and edited by World Scientific Press.

23.3 Robot Consciousness and Neuroscience

Consciousness is an important research topic in neuroscience (Rees et al. 2002; Tononi and Koch 2008; Koch et al. 2016). Many neuroscientists working on consciousness have built computational models to test their theories.

The late Nobel Prize winner Gerald Edelman, a scholar of research on biological consciousness, employed robots to validate parts of his theory. Reeke et al. (1990) discussed the Darwin series of automata (see chapter 1 for their influence in the history of cognitive robotics). They are computational systems that incorporate models of synaptic modifications, of the organization of neural cells in large assemblies, and of the integration of the actions of different cortical layers to generate the behavior of a robot according to context and its history and without the need for preprogramming the robot. Darwin I is a simple network able to recognize patterns, while Darwin II can categorize and generate associations. Darwin III is a sophisticated robot model working in a simulated environment and able to learn sensorimotor coordination, the capability of tracking objects, and the ability to reach and grasp objects and to categorize them by interacting with the environment.

Krichmar et al. (2005) discussed complex systems implemented on a real moving robot and based on computational simulations of parts of the nervous system. Darwin VII can carry out perceptual categorization and conditioned responses in simple foraging tasks, and Darwin VIII can solve the binding problem—that is, to bind the attributes of a perceived scene to form suitable coherent categories, without the need of a control system. The robot behavior emerges from the interaction of different cell assemblies without the need for preprogramming.

Stanislas Dehaene, a world-leading expert on biological consciousness, built several computational models of the neural correlates of consciousness (Dehaene et al. 2003; Zylberberg et al. 2010). In more detail, Dehaene et al. (2003) describe a computational model based on two spaces. The first space is a global neural workspace made up of distributed neurons tightly interconnected with long-range axons. The second space is a set of specialized processors related to perception, motion, memory, attention, and evaluation. Briefly, the role of the first space is to broadcast the information coming from the specialized processors belonging to the second space. The global neural workspace is tightly related to the global workspace theory (see below).

Paul Verschure (2013) analyzed the core principles of conscious states and proposed a biologically inspired architecture for perception, cognition, and action (DAC, or distributed adaptive control) to implement the core principles. Verschure claimed that the shift of research from artificial intelligence to artificial consciousness would bring more advanced machines and address the critical problem of subjective experience in humans and machines.

Recently, Dehaene et al. (2017) discussed the possibility of machine consciousness in the prestigious journal *Science*. They proposed a separation of two different information-processing aspects related to consciousness. The first aspect is related to the selection of information for global broadcasting. A second aspect is correlated to self-monitoring of these computations. The article reviewed examples of computational models inspired to machine consciousness, and it concluded with the claim that "the empirical evidence is compatible with the possibility that consciousness arises from nothing more than specific computations" (Dehaene et al. 2017, 7).

23.4 Theoretical Issues of Consciousness in Humans and Robots

A common route of investigation in robot and machine consciousness is to find a minimal set of characteristics that should be verified in an artifact before asserting whether the artifact is conscious or not.

Aleksander (1992), in the previously cited attempt, proposed five axioms that should be verified by a conscious organism. They are as follows: 1) an organism that does not learn cannot be conscious; 2) a conscious organism possesses an inner state able to represent the external world; 3) a conscious organism is able to pay attention to the contents of its internal state; 4) a conscious organism is able to generate inner states related to sequences of external inputs and to generate suitable actions; 5) the organism is able to predict external events by controlled developments of its inner state.

Aleksander and Dunmall (2003) extended this early attempt and proposed a new set of axioms for minimal consciousness in agents. These axioms are the minimal mechanisms underpinning experience. It should be noted that these authors are interested in finding a theoretical grounding for experiential consciousness in humans and artifacts. The axioms are derived from the introspective analysis of consciousness.

Let A be a generic agent in the world S. For A to be conscious of S:

- A has perceptual states that represent parts of S, corresponding to the subjective feeling that the conscious subject A is a part of, but separate from, the world S;
- A has internal states that recall elements of S or generate imagined S-like sensations, corresponding to the subjective feeling that the perception of the world S is mixed with A's past experiences;
- A can pay attention to parts of S to represent or to imagine, corresponding to the reflective feeling that A's experience of the world S is selective;
- A can control imagined state sequences to generate a plan of action, corresponding to the reflective feeling that A can think ahead of time to decide what to do;
- A has affective states able to evaluate planned operations and determine the appropriate action, corresponding to the subjective feeling that A has emotions and moods that determine its course of activities.

Aleksander and Dunmall translated these axioms in terms of mathematical constraints to be satisfied by a neural system to be considered as endowed with minimal consciousness. Aleksander (2005) proposed a schema of a cognitive architecture derived from the axioms (figure 23.1).

Selmer Bringsjord (see, e.g., Bringsjord 2007) contrasted the possibility of experiences in robots and proposed the notion of *cognitive* consciousness defined in terms of formal axioms of deontic cognitive event calculus (DCEC*; Bringsjord et al. 2018). DCEC* is a logical framework based on multisorted, quantified modal logic. It considers operators for belief, intention, knowledge, obligation, and so on. The framework allows the representation of formulae for belief and obligation. It is a family of logic in which the personal pronoun I* is based on provable theorems.

The framework provided by Bringsjord and colleagues considers the cognitive aspects of consciousness because it represents the belief about oneself and is related to a first-

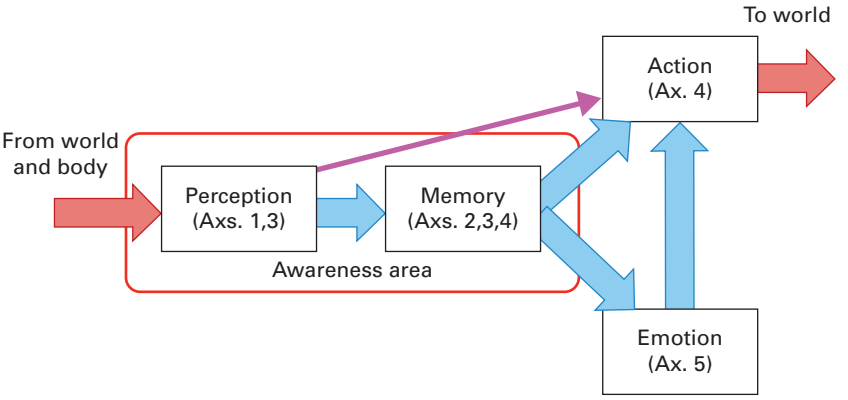

Figure 23.1
The cognitive architecture proposed by Aleksander (2005) summarizing the consciousness axioms by Aleksander and Dunmall (2003).

person representation of self-consciousness, but without considering bodily experiences. HyperSlate™ is a freely available implementation of the framework (see link in the list of additional resources).

Bringsjord et al. (2015) reported an impressive example of the framework by presenting an implementation on the NAO robot that passed the human test of self-consciousness proposed by Floridi (2005).

Giulio Tononi proposed the information integration theory (IIT) of consciousness. IIT is today the most debated scientific theory of consciousness, and many scholars actively contribute to the theory. Important outcomes also follow for robot consciousness.

The original formulation (Tononi and Sporns 2003; Tononi 2004; Tononi 2008) starts from the observation that conscious experience is differentiated because the potential repertoire of different conscious states is enormous. At the same time, conscious experience is integrated, as every conscious state is experienced as a single entity. Thus, the substrate of conscious experience must be an integrated entity able to differentiate among an enormous repertoire of different states.

The capability of a system S to differentiate among states is related to how much information can be generated by the system, and it is measured by the entropy of the system $H = -\sum p_i \log_2 p_i$, where p_i are the probabilities of the alternative outcomes of the system S.

The capability of a system S to integrate information can be measured through the effective information EI. Let us consider the system S subdivided into two partitions [A, B], and let us perturb A in order to reach the maximum entropy to outputs of A—that is, A^{Hmax}. Then, the effective information from A to B is given by $EI(A \rightarrow B) = MI(A^{Hmax}, B)$, where $MI(A, B) = H(A) + H(B) - H(AB)$ is the mutual information that measures the information shared by the source A and the target B.

The effective information EI is a measure of how the subsystem B is connected with the subsystem A. Let us consider the system S1 in figure 23.2 (*top*), where there are tight connections from A to B. Then, when A is highly perturbed, B will produce many different outputs, and $EI(A \rightarrow B)$ will be a high value.

Instead, if there are scarce or low connections between A and B, as in the case of system S2 in figure 23.2 (*bottom*), then the perturbation of A will produce scarce effects on B, and thus $EI(A \rightarrow B)$ will be a small or null value. The effective information is generally nonsymmetric, so, for a given partition, the effective information is the sum of the EI for both directions: $EI(A \leftrightarrow B) = EI(A \rightarrow B) + EI(B \rightarrow A)$. It is to be noted that if there is a partition $[A, B]$ of the system S so that $EI(A \leftrightarrow B) = 0$, then S is made up by the two independent subsystems A and B.

To measure the capability of the system to integrate information, we need to find the minimum information bipartition $MIB(S) = [A, B]$—that is, the partition $[A, B]$ of the system S for which the normalized effective information leads a minimum. $\Phi(S)$ measures the capability of the system S to integrate information, and it is the effective information given by the minimum information partition: $\Phi(S) = EI(MIB(S))$.

A subset of the system S with $\Phi > 0$ is called a *complex* when it is not included within a more substantial subset of S with a higher value of Φ. The complex of the system S with the maximum amount of $\Phi(S)$ is the *main complex*. Tononi (2004) claims that the main complex contributes to the conscious experience of S, and the measure $\Phi(S)$ grades the consciousness of the system.

Therefore, a conscious complex is a complex with a high value of $\Phi(S)$. The other parts of the systems do not contribute to the consciousness of the system. He supports his claim by analyzing different neural network models of parts of the brain and by showing that the networks with high values of $\Phi(S)$ are those typically associated with consciousness.

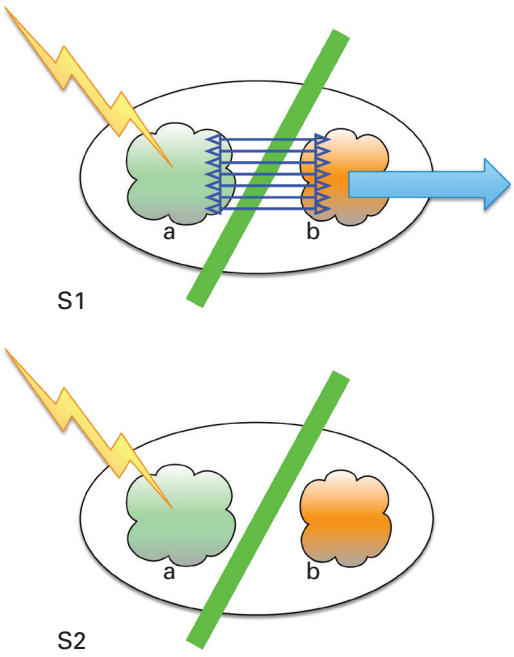

Figure 23.2
A pictorial view of a system subdivided into connected partitions A and B. *Top:* The two partitions of S1 are tightly connected, and $EI(A \rightarrow B)$ will be a high value. *Bottom:* The two partitions of S2 are barely connected, and $EI(A \rightarrow B)$ will be a low value.

Koch (2009) indicates some of the challenges of the IIT to be the unclear relationship of high values of $\Phi(S)$ with intelligence, the need for efficient algorithms for computing $\Phi(S)$ in real systems, and the need to clarify the relationships between conscious and unconscious processing.

It is to be noted that the original $\Phi(S)$ is a static measure of S; that is, it depends on the connections of the subparts of S and not on its dynamics. Balduzzi and Tononi (2008) generalize the IIT by considering the dynamics of the system. Several other extensions of IIT have been proposed in the literature; the most up-to-date version is in Oizumi et at. (2014). Tegmark (2016) investigates many variants of the original $\Phi(S)$ measure to derive exact and approximated versions that are computationally feasible to apply to real-world data.

According to IIT, experience—for example, information integration—is a fundamental quantity of nature as the mass, the charge, and the energy. Any physical system may have experiences to the extent that it can integrate information. Therefore, it could be possible in principle to build conscious artifacts by endowing them with a complex of high $\Phi(S)$. However, Kock and Tononi (2017) suggest that conventional computer architectures are unable to perform an effective integration of information, and they are unable to experience anything. A robot based on a conventional computer may be a "zombie," an entity similar to a conscious entity from its outside behavior but incapable of having real experience. Unconventional architectures, such as the neuromorphic systems, are more likely to perform the effective information integration processes happening in the brain, and therefore, they are more likely to have experience.

According to the analysis of Kock and Tononi (2008), there are many unessential ingredients for consciousness, in the sense that they have no roles in information integration. Sensory inputs and motor outputs, emotions, attention, explicit or working memory, self-reflection, and language are all capabilities that have no roles in consciousness or in robot consciousness.

Edlund et al. (2011) performed artificial life experiments to analyze the evolution of simple agents aimed to solve a maze in a simulated environment. The authors found a clear correlation between the measures of information integration and the measures of fitness of the agent, suggesting that information integration capabilities evolve and are related to the functional complexity of the agent.

23.5 Self-Consciousness in Robots and Machines

A significant topic of robot consciousness is to give a robot the capabilities of self-awareness—that is, to reflect about itself, its perceptions, and actions during its operating life. According to this approach, a computational model of the mind may be made up of a hierarchy of modules, where low-level modules are related to reactive input-outputs, and middle-level modules are related to deliberative planning and reasoning. The high-level modules are associated with self-monitor and self-reflection capabilities.

The first theoretically founded attempt to give self-reflection capabilities to an artificial reasoning system is described in the seminal paper of Weyhrauch (1980). Weyhrauch proposed the reasoning system FOL, able to perform inferences and based on a logic system and a simulation structure capable of analog representations. The system can exploit meta

representations and reflect about itself, its inferences, and its capabilities. Weyhrauch (1995) discusses the relationships between FOL and consciousness in artifacts. The original implementation of FOL is still available in LISP (see link in the list of additional resources).

An early attempt to model consciousness by considering different levels of representation is in Johnson-Laird (1983). In the well-known book on mental models, Johnson-Laird discusses consciousness as the "operating system" of the mind. Several unconscious distributed processes run in the brain, and consciousness acts as the central control system of the mind, a sort of operating system. According to this view, the content of consciousness is made up of the value parameters of the central control system.

Minsky (2006) described a multiagent system based on several interacting agents at different levels, in which the tasks of higher-level agents are self-reflection and self-consciousness (figure 23.3). In detail, Minsky proposed different levels of agents, in which each level reflects on and critiques the levels beneath.

The first levels of the system are related to agents devoted to instinctive reflexes and learned reactions. The middle level is relevant to deliberation—that is, to the prediction-planning capabilities of the system. The higher levels are related to reflection, self-reflection, and self-consciousness. In particular, the reflection level is related to the ability to criticize the deliberative techniques adopted in the previous level; the self-reflection level is associated with

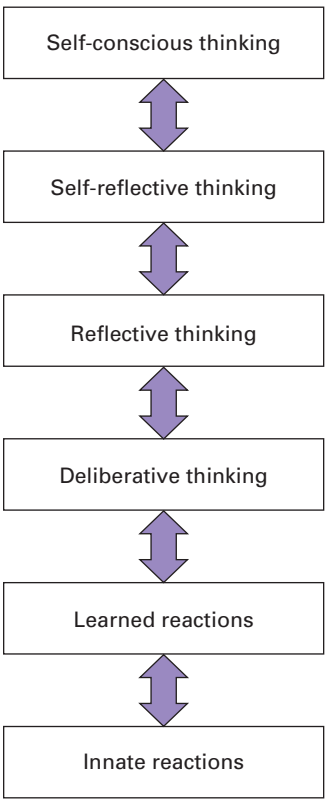

Figure 23.3
An outline of the multiagent system proposed by Minsky (2006).

the ability to generate critiques of the deficiencies and the weaknesses in the knowledge and methods employed by the system.

The higher level of the system is related to self-consciousness—that is, the ability to reflect on what others may think of the capabilities and performances of the system itself. A first attempt to implement the scheme proposed by Minsky in a simulated world was described by Singh and Minsky (2005).

Sloman and Chrisley (2003) followed a similar approach in the design of the H-CogAff architecture. H-CogAff is a framework architecture based on three primary levels related to reactive mechanisms, deliberative reasoning, and metamanagement—that is, reflective processes. The proposed framework prescribes different types of information, forms of representation, uses of data and types of mechanism for each level, and ways to put them together in the architecture. The SimAgent Toolkit is a freely available implementation in the Poplog framework.

McDermott (2001) made a distinction between *normal* access to the output of a computational module and *introspective* access to the same module. The first concerns the output related to the processing algorithms of the module. The second is related to the higher-order access within the processing of the module according to the self-model. He discussed the relationships between higher-order access and phenomenology in the line of higher-order theories of consciousness (see, e.g., Carruthers 1996).

McCarthy (1995) stressed the idea that a robot needs the ability to observe its mental states. He proposed a logic formalism to deal with aspects of self-reflection that could make robots conscious of their mental states. In detail, he presented the "mental situation calculus," an extension of the situation calculus formalism aimed at modeling introspective actions in robots.

According to the classic version of situation calculus (see, e.g., Reiter 2001), the evolution of a state of affairs in the world is modeled by a sequence of situations S_0, S_1, S_2, . . . S_n. The world changes when an instantaneous action a is performed. A new situation S_i is the result of the application of action a to the old situation S_{i-1}; then $S_i = Result(a, S_{i-1})$. In the situation calculus formalism, the truth value of a proposition p depends on the considered situation. Then the formula $Holds(p, S_i)$ means that p is true in the situation S_i.

Let us consider the situation S_i where the robot knows the proposition p—for example, the color of the object A. The formula $Holds(Know(Color(A)), S_i)$ formalizes the fact that the robot knows the color of A. The situation in which the robot infers by introspection that it does not know the color of A is formalized by the formula $Holds(Know(Not(Know(Color(A)))), S_i)$. In this case, the robot knows that it does not know the color of A. Then, because of this fact, the robot may start some actions to learn the color of A.

The mental state of the robot may evolve because of learning actions. Let us consider the previous mental situation S_i, in which the robot does not know the color of A. As an effect of teaching activities, the robot may learn the color of A. Then its mental state evolves to a new situation: $S_{i+1} = Result(Learn(Color(A))), S_i)$. The robot is in a new mental situation in which it now knows the color of A: $Holds(Knows(Color(A)), Result(Learn(Color(A))), S_i)$. Forgetting actions may be modeled similarly.

The mental situation calculus wants to capture the dynamics of self-reflection so that a robot may reason about its mental states. As emerges from the previous examples, the propositions and actions are mental, and the situations are the mental states of the robot.

In summary, the mental situation calculus is aimed at capturing the dynamic evolution of robot mental states.

Chella et al. (2008) proposed a cognitive architecture for a robot with introspective capabilities, organized in three computational areas. The *subconceptual* area is concerned with the low-level processing of perceptual data coming from the sensors. In the *linguistic* area, representation and processing are based on a logic formalism. In the conceptual area, the data coming from the subconceptual area are organized in *conceptual* categories.

Robot self-consciousness is based on the higher-order perception of the robot, in the sense that the first-order perception of the robot is the immediate perception of the environment, while higher-order perception is the perception of the inner world of the robot.

The described cognitive architecture has been tested on the board of a moving robot performing guided tours at the Archaeological Museum of Agrigento, Italy.

23.6 Global Workspace Theory

The global workspace theory (GWT) was proposed by Baars (see, e.g., Baars 1997) as the unification of different processes in the cortex. The GWT is tightly related to the global neuronal theory discussed by Dehaene et al. (2003). Baars observed that the brain could perform an enormous amount of unconscious parallel processing, while consciousness is serial and of limited capacity.

The GWT is based on assumptions that the brain is a collection of many specialized processors. Consciousness is associated with a *global workspace* whose contents "broadcast" to the processors. The processors work in parallel, and they compete to gain access to the global workspace (figure 23.4, *left*).

At some point, one processor wins the competition, and it gains access to the global workspace. Then it enters into consciousness and broadcasts to all the other processors to recruit others and to select the corresponding action (figure 23.4, *right*).

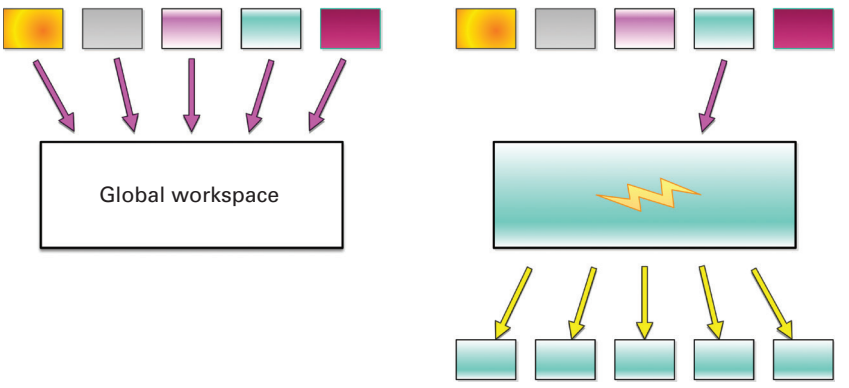

Figure 23.4
Global workspace theory. *Left:* Several unconscious processors compete to gain access to the global workspace. *Right:* The winning processor gains access to the global workspace—that is, to consciousness—and it recruits other processors.

Let us consider, for example, an agent attending an elaborate scene where there are many moving objects. According to the GWT, every moving object may be processed by an unconscious processor. All processors compete to gain access to the global workspace. Then, at some point, one processor corresponding, for example, to a ball moving toward the agent wins the competition, and it enters into consciousness. The winning processor recruits other processors to select the best action to be performed: for example, it will recruit the processors related to the motion of the arm so that the arm catches the moving ball.

Contexts shape conscious contents, and they constrain the competition of unconscious processors. Therefore, a coalition of processors may be expedited to gain access in a particular context and to recruit other processors. For example, a context related to a specific emotion may assist processors in achieving consciousness instead of other processors.

The GWT is a framework theory, and several cognitive architectures inspired by the GWT have been proposed in the literature. The main cognitive architecture is LIDA (Learning Intelligent Distributed Agent), developed by Stan Franklin and colleagues over the years (see, e.g., Franklin et al. 2014; see also chapter 10 for a general discussion of cognitive architectures).

Baars and Franklin (2009) reported on the relationships between LIDA and the GWT. An initial version of LIDA, named IDA, was built by Franklin (2003) as a dispatching system for the US Navy. The goal of IDA was to assign sailors to new billets at the end of their tours of duty. These assignments were performed by detailers, and IDA completely automated the roles of detailers. Interaction with sailors was performed by email in natural language, and IDA was able to negotiate the new billets with sailors and to write orders to them.

An overview of LIDA is shown in figure 23.5. Several processors based on different technologies were implemented in the architecture, such as neural networks, sparse distributed memories, schema mechanisms, behavior networks, and subsumption architectures. LIDA performs several aspects of the GWT, like perception, attention, episodic and declarative memories, the global workspace, and the selection of actions.

The cognitive cycle of LIDA is based on the following steps:

– The system perceives an entity, giving rise to a percept.

– The percept is sent to a preconscious buffer, where the percept gives rise to local associations.

– The percept competes for consciousness.

– If the percept wins the competition, then it broadcasts to all the other processors to recruit for resources.

– An action is selected according to the goal context hierarchy.

– Once the action is selected, then the action is executed, and the cognitive cycle restarts.

The chosen action may be performed immediately, or it may be sent back to the perceptual system for further examinations.

The LIDA architecture presents learning capabilities through the feedback generated by the global workspace. The feedback signals are sent to the unconscious modules, and they provide the basis of the reinforcement- and associative-learning processes of the architecture. The Lidapy framework is a freely available recent implementation of LIDA in Python (see link in additional resource list).

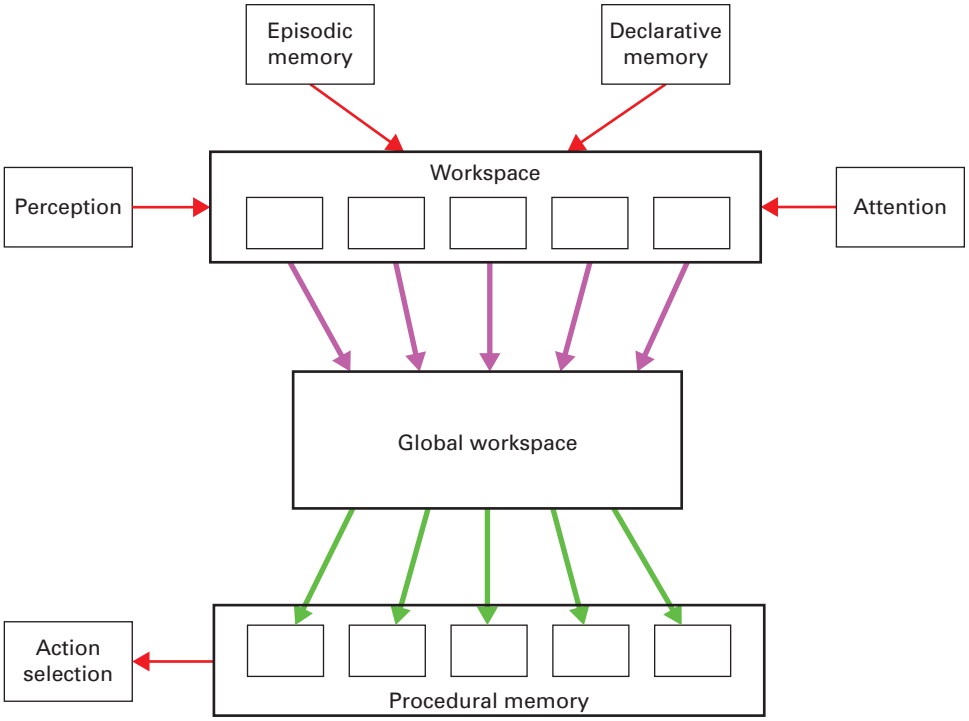

Figure 23.5
An overview of the LIDA cognitive architecture.

The LIDA architecture has proved to fit a body of empirical evidence concerning consciousness. Notably, a version of LIDA (Madl et al. 2011) implementing the Allport (1968) test modeling the phenomenal simultaneity of stimuli obtained time frames comparable to human subjects. Ramamurthy and Franklin (2009) discuss the general problems of conscious experiences and functional consciousness in the framework of LIDA.

Other cognitive architectures inspired by the GWT have been proposed in the literature. Shanahan (2006) discussed a cognitive architecture for a robot that extends the GWT by considering a cognitive cycle made up of an inner and an outer loop. The outer loop is similar to the cycle previously discussed in LIDA, while the role of the inner loop is to simulate the interaction with the environment internally. The internal simulation facilitates anticipation and planning in the architecture: the robot may internally simulate the effects of the actions before choosing the current course of activities.

Arrabales et al. (2009) discussed CERA-CRANIUM, a cognitive architecture based on GWT that controls a video game character. The architecture performed well in the BotPrize competition (Hingston 2009), a kind of Turing test (see below) in which autonomous bots have to convince a jury that they are human controlled. Notably, the CERA-CRANIUM bot won the award for the most humanlike bot at the 2010 competition. The software code of the bot is freely available (see link in the additional resource list).

Haikonen (see, e.g., Haikonen 2019), starting from engineering principles, designed the HCA, or Haikonen cognitive architecture, which presents contact points with the GWT.

The HCA is at the basis of the operating robot XCR-1, where many modules are implemented, including the auditory module, the visual module, and the emotional module. The modules send broadcast signals and compete in a winner-takes-all fashion to control the robot, similar to GWT. XCR-1 presents many aspects of machine consciousness: the robot can self-talk, respond to visual stimuli, and "feel" pain and emotions, among other functionalities.

23.7 The Internal Model Hypothesis

The internal model hypothesis states that an agent, to act in an intelligent and meaningful way, operates via an internal model of itself and the external world. The internal model allows the agent the capability to simulate its actions and evaluate its outcomes before doing them in the external environment. In this way, the agent can generate expectations about the course of events in the world and on the outcomes of its actions.

The internal model hypothesis is inspired by the "small-scale model" of reality discussed by Craik (1943). Dennett (1996) discusses "Popperian" creatures—that is, creatures able to generate theories about the external world and simulate experiments in their internal environment.

The proposal of an internal model acting as a simulation structure in a robot is not new: robot architectures have been proposed in the literature that present forms of an internal model of themselves and the external environment. Early examples have been provided by Mel (1990), Stein (1994), and Payton (1990), among others.

According to Hesslow (2002), the internal model hypothesis allows the brain to simulate actions, to simulate perceptions, and to generate anticipation about future events. Hesslow claims that conscious thoughts are based on these simulations. As the simulation of perception is related to the internally generated sensory inputs resembling the perception of the external world, it would be accompanied by the experience of the internal model of the world.

In brief, the internal model hypothesis states that consciousness arises from interaction between the internal model of the agent and the internal model of the world. Let us consider an agent interacting with the external world (figure 23.6, *top*).

Let us now consider the internal model of the agent, including the model of the agent and the external world (figure 23.6, *bottom*). According to the internal model hypothesis, consciousness arises not from the interaction of the agent with the external world but instead from the interaction of the internal model of the agent with the internal model of the external world. Susan Blackmore (1986) states that "being conscious is simply what it is like being a representation of the world" (163).

Figure 23.7, inspired by Grush (2004) and Gerdes and Happee (1994), describes the general framework of the internal model. A similar structure has been presented by Gray (2006). The robot has an internal model of itself and the external environment, allowing it to simulate its interactions with the external world. The controller sends the control signal at the same time to the real robot moving in the external world and to the inner model of the robot moving in the inner environment. Again, according to the internal model hypothesis, robot consciousness arises in the interaction of the internal model of the robot with the internal model of the situation.

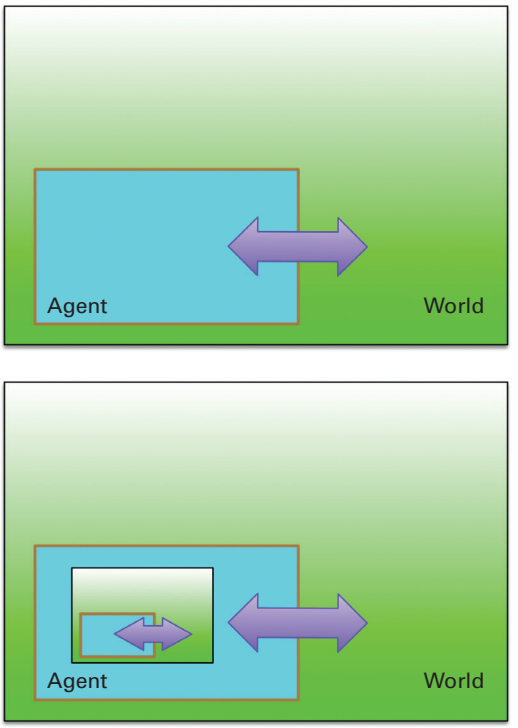

Figure 23.6
The internal model hypothesis. *Top:* The agent interacting with the external world. *Bottom:* The agent with an internal model of itself interacting with an internal model of the external world.

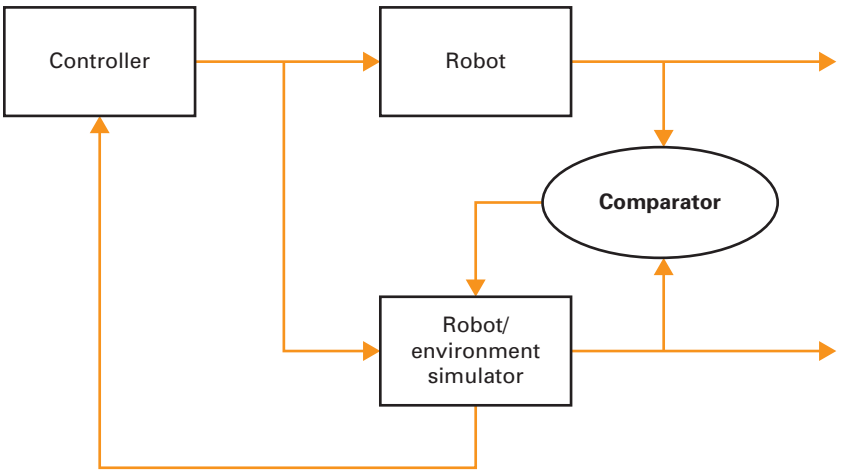

Figure 23.7
A general framework of the internal model hypothesis for robot consciousness.

A robot implementation inspired by the internal model hypothesis is EcceRobot, developed by Holland and colleagues (Holland 2007; Holland et al. 2007). EcceRobot is an anthropometric robot with a humanlike body. The robot has an internal simulator of itself and the environment that is able to represent in three dimensions (3D) the robot and the environment. The internal 3D simulation is employed to teach suitable neural networks how to control the motors of the robot.

Bongard et al. (2006) describe a "starfish" robot, a four-legged robot that generates a 3D model of itself by trial and error using suitable genetic algorithms. The robot uses the actuation-sensation relationship to infer an internal model of its body, and then it uses this model to learn locomotion. The robot is resilient: in case of damage—for example, a broken leg—the robot can generate a new model of its body and learn locomotion again with its current damaged body. A similar approach was described by Cully et al. (2015).

Chella and Macaluso (2009) discussed the robot CiceRobot, which was able to offer guided tours in an indoor and outdoor museum and was based on the internal model hypothesis. The architecture was instantiated on a wheeled robot for indoor and outdoor use. Currently, it is instantiated on a Pepper robot. The robot is a case study of many capabilities associated with the functional aspects of consciousness: to build and to maintain an internal model of the environment and itself, to pay attention to the relevant entities in the environment, to integrate information from different sources and different parts of the same source, to generate expectations about the possible events in the environment, to self-monitor, to simulate emotional states, and to process information by making it globally available to the robot.

The primary outcome of the case study was the acceptancy and transparency of the autonomous behavior of the robot in an environment populated by untrained users as museum tourists.

23.8 Tests for Robot Consciousness

People are concerned that current robot systems might already be conscious, so a substantial amount of research has been conducted on how a robot system can be tested for consciousness. An extended review of proposed criteria for consciousness in machines and robots is discussed by Elamrani and Yampolskiy (2019).

Many tests are based on the famous Turing (1950) test of imitation, in which a human interrogates an entity by teletype and decides whether they are examining a human or a machine that imitates human responses.

Sloman (2010) proposed the Robot Philosopher Test, a variant of the Turing test in which the arguments of discussion between the human tester and the tested entity are the philosophical theories of consciousness and experience.

Schneider and Turner (2017; see also Turner and Schneider 2019) proposed the Artificial Consciousness test (ACT), another variant of the Turing test in which the questions to be posed are focused on the quality of the inner experience of the entity under examination. The entity must be isolated from the external world to avoid the risk that a smart machine may retrieve the correct answers from the internet.

Harnad (1991) extended the Turing test by proposing the Total Turing test, in which a robot—that is, an embodied entity—should imitate the whole of human behavior in different situations.

Another source of inspiration for consciousness tests is the mirror test for primates (Gallup 1970; Gallup et al. 2002). In this case, a robot should recognize and describe itself and its movements by looking through a mirror, even in the presence of other robots and distractors. See Gold and Scassellati (2005), Chella et al. (2003), Suzuki et al. (2005), and Haikonen (2007) for examples of robot implementations of the mirror test.

Consciousness in robots and machines can be assessed by measuring specific features ascribed to consciousness, like the ability to presents forms of creativity—that is, to produce something new and unexpected. Bringsjord et al. (2001) presented the Lovelace Test, named after Ada Lovelace, while Chella and Manzotti (2012) discussed how a conscious robot should be able to improvise jazz in a jazz ensemble.

A related approach is to consider the capability of the conscious robot to generate a genuine inner speech, as proposed by Haikonen (2007). Inner speech is considered tightly related to self-consciousness (Morin 2005). Steels (2003), Clowes (2007), Arrabales (2012), and Chella et al. (2020) demonstrate examples of robots presenting forms of inner speech.

Another approach for testing machine consciousness is to apply the algorithmic theories proposed for human and robot consciousness, such as the previously described set of axioms by Aleksander and Dunmall or the $\Phi(S)$ measure derived from the information integration theory.

Gamez (2010) implemented SpikeStream, a freely available neural network simulator able to measure the $\Phi(S)$ of different kinds of networks (see the link to the system in the list of additional resources). In detail, Gamez applied $\Phi(S)$ to analyze the neural networks at the basis of SIMNOS, a simulation of EcceRobot.

Iklé et al. (2019) followed a similar approach to measure $\Phi(S)$ in the cognitive system controlling the robot Sophia when the robot was reading and when it was conversing. Seth et al. (2006) and Gamez and Aleksander (2009) proposed methods for designing suitable neural networks presenting high values of the measure $\Phi(S)$.

An interesting approach to assess consciousness in robots and machines was proposed by Arrabales et al. (2010a). They discussed ConsScale, a scale of consciousness in artificial agents that scores from −1 and 0 (the disembodied and isolated agent) to 11 (the superconscious agent).

ConsScale considers a generic characterization of an artificial agent to comprise a body, a set of sensors, a set of actuators, a set of software routines, types of memories, and an external environment where the agent operates.

ConsScale assigns a level of consciousness according to the architectural complexity of the agent and to the behaviors of the agent. At the low level of ConsScale are reactive agents based on a direct link between sensors and actuators. At the intermediate levels are the agents able to adapt themselves, to pay attention, to generate plans, and to have emotions. At the higher level of the scale are the self-conscious agents, the empathic agents, and the social agents. At the top level is the humanlike agent, which can pass the Turing test, and the superconscious agent, able to manage several streams of consciousness. The ConsScale calculator is freely available (see the link in the list of additional resources).

Arrabales et al. (2010b) tested ConsScale by assessing some cognitive architectures such as CERA-CRANIUM, CRONOS (an implementation of EcceRobot), LIDA, and a version of the HCA. According to the assessment by ConsScale, the HCA and LIDA

received the highest score because they were successful at the emotional level—that is, at an intermediate level of consciousness. No architectures entered the higher levels.

23.9 Conclusion

Chella and Manzotti (2009) wrote a manifesto for robot consciousness in which they discussed some of the main challenges in the field. Notwithstanding the progress in this field, as seen in the numerous machine consciousness theories presented above, the challenges from this manifesto are still valid today. They include the role of embodiment and situatedness in machine consciousness, the roles of emotion and motivation, the difficulties in achieving information integration, the concept of time for robot consciousness, the question of free will for robots, and finally, the issue of robot experience.

The possible advent of conscious robots would lead to ethical concerns as well as issues related to the social integration of such robots. Bryson (2012, 2018) discussed in detail the risks of our moral obligations toward self-conscious systems. According to Bryson (2018, 15), "While constructing AI systems as either moral agents or patients is possible, neither is desirable."

According to Gunkel (2012), if an entity has subjective experiences and is capable of suffering, then it should be treated as a person. These arguments may force us to review our fundamental definitions of the concept of person. If we assert that a robot system is conscious, then the moral responsibility of the system for its actions must be recognized. On the other hand, we may have to concede moral rights to conscious robots, such as the right to not be switched off.

In summary, robot consciousness is a research field that not only offers outstanding opportunities but brings ethical risks that cannot be undervalued.

Additional Reading and Resources

- This collection of classic papers on machine and robot consciousness is a valuable academic reference in the field: Chella, A., and R. Manzotti, eds. 2007b. *Artificial Consciousness*. Exeter, UK: Imprint Academic.
- This book is an introduction to robot consciousness from the perspectives of philosophy, cognitive science, and computer science, written by a founding father of the discipline: Aleksander, I. 2015. *Impossible Minds: My Neurons, My Consciousness*. Rev. ed. Singapore: World Scientific.
- This freely available e-book is a collection of papers that cover the most recent research trends of consciousness in robots and AI systems: https://www.frontiersin.org/research-topics/5781/consciousness-in-humanoid-robots. Chella, A., A. Cangelosi, G. Metta, and S. Bringsjord, eds. 2019. *Consciousness in Humanoid Robots*. Lausanne: Frontiers Media. doi:10.3389/978-2-88945-866-0.
- This new journal, with a freely available inaugural issue, presents the latest works in the field of consciousness in robotics and AI: https://www.worldscientific.com/worldscinet/jaic.
- HyperSlate™ logical framework by Bringsjord: https://rpi.logicamodernapproach.com/.

- Reasoning system FOL by Weyhrauch: https://github.com/getfol/GETFOL.
- The SimAgent Toolkit by Aaron Sloman: https://www.cs.bham.ac.uk/research/projects/poplog/packages/simagent.html.
- The LIDA framework: https://github.com/CognitiveComputingResearchGroup/lidapy-framework.
- The CERA-CRANIUM bot: https://github.com/raul-arrabales/CCbot4.
- The SpikeStream simulator by Gamez: http://spikestream.sourceforge.net/.
- The ConsScale consciousness calculator by Arrabales: https://www.conscious-robots.com/consscale/calc_30.html.

Note

1. Swartz Foundation, *Final Report of the Workshop Can a Machine Be Conscious*, 2001, http://www.theswartzfoundation.org/abstracts/2001_summary.asp.

References

Aleksander, Igor. 1992. "Capturing Consciousness in Neural Systems." In *Artificial Neural Networks 2, Proceedings of ICANN 1992 Conference*, 17–22. Amsterdam: Elsevier.

Aleksander, Igor. 2005. *The World in My Mind, My Mind in the World*. Exeter, UK: Imprint Academic.

Aleksander, Igor. 2015. *Impossible Minds: My Neurons, My Consciousness*. Revised ed. Singapore: World Scientific.

Aleksander, Igor, and Barry Dunmall. 2003. "Axioms and Tests for the Presence of Minimal Consciousness in Agents." *Journal of Consciousness Studies* 10 (4–5): 7–18.

Allport, David A. 1968. "Phenomenal Simultaneity and the Perceptual Moment Hypothesis." *British Journal of Psychology* 59:395–406.

Arrabales, Raúl. 2012. "Inner Speech Generation in a Video Game Non-player Character: From Explanation to Self?" *International Journal of Machine Consciousness* 4 (2): 367–381.

Arrabales, Raúl, Agapito Ledezma, and Araceli Sanchis. 2009. "Towards Conscious-Like Behavior in Computer Game Characters." In *Proceedings of the IEEE International Conference on Computational Intelligence and Games*, 217–224. Piscataway, NJ: IEEE Press.

Arrabales, Raúl, Agapito Ledezma, and Araceli Sanchis. 2010a. "ConsScale: A Pragmatic Scale for Measuring the Level of Consciousness in Artificial Agents." *Journal of Consciousness Studies* 17 (3–4): 131–164.

Arrabales, Raúl, Agapito Ledezma, and Araceli Sanchis. 2010b. "The Cognitive Development of Machine Consciousness Implementations." *International Journal of Machine Consciousness* 2 (2): 213–225.

Baars, Bernard J. 1997. *In the Theater of Consciousness: The Workspace of the Mind*. Oxford: Oxford University Press.

Baars, Bernard J., and Stan Franklin. 2009. "Consciousness Is Computational: The LIDA Model of Global Workspace Theory." *International Journal of Machine Consciousness* 1 (1): 23–32.

Balduzzi, David, and Giulio Tononi. 2008. "Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework." *PLoS Computational Biology* 4 (6): e1000091. https://doi.org/10.1371/journal.pcbi.1000091

Blackmore, Susan. 1986. "What It's Like to Be a Mental Model." In *Research in Parapsychology*, edited by D. Weiner and D. Radin, 163–164. Metuchen, NJ: Scarecrow.

Blackmore, Susan, and Emily T. Troscianko. 2018. *Consciousness—an Introduction*. London: Routledge.

Bongard, Josh, Victor Zykov, and Hod Lipson. 2006. "Resilient Machines through Continuous Self-Modeling." *Science* 314:1118–1123.

Bringsjord, Selmer. 2007. "Offer: One Billion Dollars for a Conscious Robot. If You're Honest, You Must Decline." *Journal of Consciousness Studies* 14 (7): 28–43.

Bringsjord, Selmer, Paul Bello, and David Ferrucci. 2001. "Creativity, the Turing Test, and the (Better) Lovelace Test." *Minds and Machines* 11:3–27.

Bringsjord, Selmer, Paul Bello, and Naveen Sundar Govindarajulu. 2018. "Toward Axiomatizing Consciousness." In *The Bloomsbury Companion to the Philosophy of Consciousness*, edited by D. Jacquette, 289–324. London: Bloomsbury Academic.

Bringsjord, Selmer, John Licato, Naveen Sundar Govindarajulu, Rikhiya Ghosh, and Atriya Sen. 2015. "Real Robots That Pass Human Tests of Self-Consciousness." In *24th IEEE International Symposium on Robot and Human Interactive Communication*, 498–504. Piscataway, NJ: IEEE Press.

Bryson, Joanna. 2012. "A Role for Consciousness in Action Selection." *International Journal of Machine Consciousness* 4 (2): 471–482.

Bryson, Joanna. 2018. "Patience Is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics." *Ethics and Information Technology* 20:15–26.

Carruthers, Peter. 1996. *Language, Thought and Consciousness: An Essay in Philosophical Psychology*. Cambridge: Cambridge University Press.

Chalmers, David J. 1995. "Facing Up to the Problem of Consciousness." *Journal of Consciousness Studies* 2 (3): 200–219.

Chella, Antonio, Angelo Cangelosi, Giorgio Metta, and Selmer Bringsjord. 2019. "Editorial: Consciousness in Humanoid Robots." *Frontiers in Robotics and AI* 6:17. https://doi.org/10.3389/frobt.2019.00017.

Chella, Antonio, Marcello Frixione, and Salvatore Gaglio. 2003. "Anchoring Symbols to Conceptual Spaces: The Case of Dynamic Scenarios." *Robotics and Autonomous Systems* 43:175–188.

Chella, Antonio, Marcello Frixione, and Salvatore Gaglio. 2008. "A Cognitive Architecture for Robot Self-Consciousness." *Artificial Intelligence in Medicine* 44:147–154.

Chella, Antonio, and Irene Macaluso. 2009. "The Perception Loop in Cicerobot, a Museum Guide Robot." *Neurocomputing* 72:760–766.

Chella, Antonio, and Riccardo Manzotti, eds. 2007a. *AI and Consciousness: Theoretical Foundations and Current Approaches, Papers from the 2007 AAAI Fall Symposium*. Menlo Park, CA: AAAI Press.

Chella, Antonio, and Riccardo Manzotti, eds. 2007b. *Artificial Consciousness*. Exeter, UK: Imprint Academic.

Chella, Antonio, and Riccardo Manzotti. 2009. "Machine Consciousness: A Manifesto for Robotics." *International Journal of Machine Consciousness* 1 (1): 33–51.

Chella, Antonio, and Riccardo Manzotti. 2012. "Jazz and Machine Consciousness: Towards a New Turing Test." In *Revisiting Turing and His Test: Comprehensiveness, Qualia, and the Real World*, edited by Vincent C. Müller and Aladdin Ayesh, 49–53. Birmingham, UK: AISB/IACAP.

Chella, Antonio, Arianna Pipitone, Alain Morin, and Famira Racy. 2020. "Developing Self-Awareness in Robots via Inner Speech." *Frontiers in Robotics and AI* 7:16. https://doi.org/10.3389/frobt.2020.00016

Clowes, Robert. 2007. "A Self-Regulation Model of Inner Speech and Its Role in the Organisation of Human Conscious Experience." *Journal of Consciousness Studies* 14 (7): 59–71.

Clowes, Robert, Steve Torrance, and Ron Chrisley. 2007. "Machine Consciousness: Embodiment and Imagination." *Journal of Consciousness Studies* 14 (7): 7–14.

Craik, Kenneth J. W. 1943. *The Nature of Explanation*. Cambridge: Cambridge University Press.

Cully, Antoine, Jeff Clune, Danesh Tarapore, and Jean-Baptiste Mouret. 2015. "Robots That Can Adapt Like Animals." *Nature* 521:503–507.

Dehaene, Stanislas, Hakwan Lau, and Sid Kouider. 2017. "What Is Consciousness, and Could Machines Have It?" *Science* 358:486–492.

Dehaene, Stanislas, Claire Sergent, and Jean-Pierre Changeux. 2003. "A Neuronal Network Model Linking Subjective Reports and Objective Physiological Data during Conscious Perception." *Proceedings of the National Academy of Sciences USA* 100 (14): 8520–8525.

Dennett, Daniel. 1996. *Darwin's Dangerous Idea*. New York: Simon and Schuster.

Edlund, Jeffrey A., Nicolas Chaumont, Arend Hintze, Christof Koch, Giulio Tononi, and Christoph Adami. 2011. "Integrated Information Increases with Fitness in the Evolution of Animats." *PLoS Computational Biology* 7 (10): e1002236.

Edelman, Gerald M., George N. Reeke, W. Einar Gall, Giulio Tononi, Douglas Williams, and Olaf Sporns. 1992. "Synthetic Neural Modeling Applied to a Real-World Artifact." *Proceedings of the National Academy of Sciences USA* 89:7267–7271.

Elamrani, Aida, and Roman V. Yampolskiy. 2019. "Reviewing Tests for Machine Consciousness." *Journal of Consciousness Studies* 26 (5–6): 35–64.

Floridi, Luciano. 2005. "Consciousness, Agents and the Knowledge Game." *Mind and Machines* 15:415–444.

Franklin, Stan. 2003. "IDA—a Conscious Artifact?" *Journal of Consciousness Studies* 10 (4–5): 47–66.

Franklin, Stan, Tamas Madl, Sidney D'Mello, and Javier Snaider. 2014. "LIDA: A Systems-Level Architecture for Cognition, Emotion, and Learning." *IEEE Transactions on Autonomous Mental Development* 6 (1): 19–41.

Gallup Jr., Gordon G. 1970. "Chimpanzees: Self-Recognition." *Science* 167 (3914): 86–87.

Gallup Jr., Gordon G., James R. Anderson, and Daniel J. Shillito. 2002. "The Mirror Test." In *The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition*, edited by M. Bekoff, C. Allen, and G. Burghardt, 325–333. Cambridge, MA: MIT Press.

Gamez, David. 2010. "Information Integration Based Predictions about the Conscious States of a Spiking Neural Network." *Consciousness and Cognition* 19 (1): 294–310.

Gamez, David, and Igor Aleksander. 2009. "Taking a Mental Stance towards Artificial Systems." In *Biologically Inspired Cognitive Architectures: Papers from the 2009 AAAI Fall Symposium*. Menlo Park, CA: AAAI Press.

Gamez, David, Zafeirios Fountas, and Andreas K. Fidjeland. 2013. "A Neurally-Controlled Computer Game Avatar with Human-Like Behavior." *IEEE Transactions on Computational Intelligence and AI in Games* 5 (1): 1–14.

Gerdes, V. G. J., and Riender Happee. 1994. "The Use of an Internal Representation in Fast Goal-Directed Movements: A Modeling Approach." *Biological Cybernetics* 70:513–524.

Gold, Kevin, and Brian Scassellati. 2005. "Learning about the Self and Others through Contingency." In *Developmental Robotics: Papers from the 2005 AAAI Spring Symposium*. Menlo Park, CA: AAAI Press.

Gray, Jeffrey A. 2006. *Consciousness: Creeping Up on the Hard Problem*. Oxford: Oxford University Press.

Grush, Rick. 2004. "The Emulator Theory of Representation: Motor Control, Imagery and Perception." *Behavioral and Brain Sciences* 27:377–442.

Gunkel, David J. 2012. *The Machine Question*. Cambridge, MA: MIT Press.

Haikonen, Pentti O. 2007a. "Reflections of Consciousness: The Mirror Test." In *AI and Consciousness: Theoretical Foundations and Current Approaches: Papers from the 2007 AAAI Fall Symposium*, 67–71. Menlo Park, CA: AAAI Press.

Haikonen, Pentti O. 2007b. *Robot Brains: Circuits and Systems for Conscious Machines*. Hoboken, NJ: John Wiley and Sons.

Haikonen, Pentti O. 2019. *Consciousness and Robot Sentience*. 2nd ed. Singapore: World Scientific Press.

Harnad, Stevan. 1991. "Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem." *Minds and Machines* 1 (1): 43–54.

Hesslow, Germund. 2002. "Conscious Thought as Simulation of Behavior and Perception." *Trends in Cognitive Sciences* 6 (6): 242–247.

Hingston, Philip. 2009. "The 2K BotPrize." In *Proceedings of IEEE International Conference on Computational Intelligence and Games*, 1–1. Piscataway, NJ: IEEE Press.

Holland, Owen. 2003a. *Machine Consciousness*. Exeter, UK: Imprint Academic.

Holland, Owen. 2003b. "Robots with Internal Models—a Route to Machine Consciousness?" *Journal of Consciousness Studies* 10 (4–5): 77–109.

Holland, Owen. 2007. "A Strongly Embodied Approach to Machine Consciousness." *Journal of Consciousness Studies* 14 (7): 97–110.

Holland, Owen, Rob Knight, and Richard Newcombe. 2007. "A Robot-Based Approach to Machine Consciousness." In *Artificial Consciousness*, edited by A. Chella and R. Manzotti. Exeter, UK: Imprint Academic.

Iklé, Matthew, Ben Goertzel, Misgana Bayetta, George Sellman, Comfort Cover, Jennifer Allgeier, Robert Smith, et al. 2019. "Using Tononi Phi to Measure Consciousness of a Cognitive System While Reading and Conversing." In Vol. 2287, *Towards Conscious AI Systems: Papers of the AAAI 2019 Spring Symposium*. Palo Alto, CA: CEUR Workshop Proceedings. http://ceur-ws.org/vol-2287/paper20.pdf.

Johnson-Laird, Philip N. 1983. *Mental Models: Towards a Cognitive Science of Language, Inference and Consciousness*. Cambridge: Cambridge University Press.

Koch, Christof. 2009. "A Theory of Consciousness." *Scientific American Mind*, July/August, 16–19.

Koch, Christof, Marcello Massimini, Melanie Boly, and Giulio Tononi. 2016. "Neural Correlates of Consciousness: Progress and Problems." *Nature Reviews Neuroscience* 17 (5): 307–323.

Koch, Christof, and Giulio Tononi. 2008. "Can Machines Be Conscious?" *IEEE Spectrum*, 47–51.

Koch, Christof, and Giulio Tononi. 2017. "Can We Quantify Machine Consciousness?" *IEEE Spectrum*, 65–69.

Krichmar, Jeffrey L., Douglas A. Nitz, Joseph A. Gally, and Gerald M. Edelman. 2005. "Characterizing Functional Hippocampal Pathways in a Brain-Based Device as It Solves a Spatial Memory Task." *Proceedings of the National Academy of Sciences of the USA* 102 (6): 2111–2116.

Madl, Tamas, Bernard J. Baars, and Stan Franklin. 2011. "The Timing of the Cognitive Cycle." *PLoS One* 6 (4): e14803. https://doi.org/10.1371/journal.pone.0014803.

Mel, Bartlett. 1990. *Connectionist Robot Motion Planning*. Cambridge, MA: Academic Press.

McCarthy, John. 1995. "Making Robots Conscious of Their Mental States." *Machine Intelligence* 15:3–17. http://jmc.stanford.edu/articles/consciousness.html.

McDermott, Drew. 2001. *Mind and Mechanisms*. Cambridge, MA: MIT Press.

Minsky, Marvin. 2006. *The Emotion Machine*. New York: Simon and Schuster.

Morin, Alain. 2005. "Possible Links between Self-Awareness and Inner Speech." *Journal of Consciousness Studies* 12 (4–5): 115–134.

Nagel, Thomas. 1974. "What Is It Like to Be a Bat?" *Philosophical Review* 83 (4): 435–450.

Oizumi, Masafumi, Larissa Albantakis, and Giulio Tononi. 2014. "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0." *PLoS Computational Biology* 10 (5): e1003588. https://doi.org/10.1371/journal.pcbi.1003588.

O'Regan, J. Kevin, and Alva Noë. 2001. "A Sensorimotor Account of Vision and Visual Consciousness." *Behavioral and Brain Sciences* 24:939–973.

Payton, David W. 1990. "Internalized Plans: A Representation for Action Resources." *Robotics Autonomous Systems* 6 (1): 89–103.

Ramamurthy, Uma, and Stan Franklin. 2009. "Resilient Architectures to Facilitate Both Functional Consciousness and Phenomenal Consciousness in Machines." *International Journal of Machine Consciousness* 1 (2): 243–253.

Reeke, George N., Olaf Sporns, and Gerald M. Edelman. 1990. "Synthetic Neural Modeling: The 'Darwin' Series of Recognition Automata." *Proceedings of the IEEE* 78 (9): 1498–1530.

Rees, Geraint, Gabriel Kreiman, and Christof Koch. 2002. "Neural Correlates of Consciousness in Humans." *Nature Reviews Neuroscience* 3 (4): 261–270.

Reggia, James A. 2013. "The Rise of Machine Consciousness: Studying Consciousness with Computational Models." *Neural Networks* 44:112–131.

Reggia, James A., Garrett E. Katz, and Gregory P. Davis. 2018. "Humanoid Cognitive Robots That Learn by Imitating: Implications for Consciousness Studies." *Frontiers in Robotics and AI* 5:1. https://doi.org/10.3389/frobt.2018.00001.

Reiter, Raymond. 2001. *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*. Cambridge, MA: MIT Press.

Rushby, John, and Daniel Sanchez. 2018. *Technology and Consciousness Workshop Report*. SRI International. http://www.csl.sri.com/users/rushby/papers/techconscwks2017.pdf.

Schmidhuber, Juergen. 1992. "Learning Complex, Extended Sequences Using the Principle of History Compression." *Neural Computation* 4 (2): 234–242.

Schneider, Susan, and Edwin Turner. 2017. "Is Anyone Home? A Way to Find Out if AI Has Become Self-Aware." *Scientific American* (blog). https://blogs.scientificamerican.com/observations/is-anyone-home-a-way-to-find-out-if-ai-has-become-self-aware/.

Searle, John R. 2000. "Consciousness." *Annual Review of Neuroscience* 23:557–578.

Seth, Anil K., Eugene Izhikevich, George N. Reeke, and Gerald M. Edelman. 2006. "Theories and Measures of Consciousness: An Extended Framework." *Proceedings of the National Academy of Sciences of the USA* 103 (28): 10799–10804.

Shanahan, Murray P. 2006. "A Cognitive Architecture That Combines Internal Simulation with a Global Workspace." *Consciousness and Cognition* 15:433–449.

Singh, Push, and Marvin Minsky. 2005. "An Architecture for Cognitive Diversity." In *Visions of Mind*, edited by D. Davis, 312–331. London: Idea Group.

Sloman, Aaron. 2010. "Machine Consciousness: Response to Commentaries." *International Journal of Machine Consciousness* 2 (1): 75–116.

Sloman, Aaron, and Ron Chrisley. 2003. "Virtual Machines and Consciousness." *Journal of Consciousness Studies* 10 (4–5): 133–172.

Steels, Luc. 2003. "Language Re-entrance and the 'Inner Voice.'" *Journal of Consciousness Studies* 10 (4–5): 173–185.

Stein, Lynn A. 1994. "Imagination and Situated Cognition." *Journal of Experimental and Theoretical Artificial Intelligence* 6 (4): 303–407.

Suzuki, Tohru, Keita Inaba, and Junichi Takeno. 2005. "Conscious Robot That Distinguishes between Self and Others and Implements Imitation Behavior." In *International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA/AIE) 2005*, edited by M. Ali and F. Esposito, 101–110. LNAI 3533. Heidelberg: Springer.

Tegmark, Max. 2016. "Improved Measures of Integrated Information." *PLoS Computational Biology* 12 (11): e1005123. https://doi.org/10.1371/journal.pcbi.1005123.

Tononi, Giulio. 2004. "An Information Integration Theory of Consciousness." *BMC Neuroscience* 5:42. https://doi.org/10.1186/1471-2202-5-42.

Tononi, Giulio. 2008. "Consciousness as Integrated Information: A Provisional Manifesto." *Biology Bulletin* 215:216–242.

Tononi, Giulio, Melanie Boly, Marcello Massimini, and Christof Koch. 2016. "Integrated Information Theory: From Consciousness to Its Physical Substrate." *Nature Reviews Neuroscience* 17 (7): 450–461.

Tononi, Giulio, and Cristof Koch. 2008. "The Neural Correlates of Consciousness: An Update." *Annals of the New York Academy of Sciences* 1124:239–261.

Tononi, Giulio, and Olaf Sporns. 2003. "Measuring Information Integration." *BMC Neuroscience* 4:31. https://doi.org/10.1186/1471-2202-4-31.

Turing, Alan. 1950. "Computing Machinery and Intelligence." *Mind* 59 (236): 433–460.

Turner, Edwin, and Susan Schneider. 2019. "Testing for Synthetic Consciousness: The ACT, the Chip Test, the Unintegrated Chip Test, and the Extended Chip Test." In Vol. 2287, *Towards Conscious AI Systems: Papers of the AAAI 2019 Spring Symposium*. Palo Alto, CA: CEUR Workshop Proceedings. http://ceur-ws.org/vol-2287/short2.pdf.

Verschure, Paul. 2013. "From the Mirage of Intelligence to a Science and Engineering of Consciousness." *IEEE Intelligent Systems*, September/October, 7–10.

Vimal, Ram L. P. 2009. "Meaning Attributed to the Term 'Consciousness'—an Overview." *Journal of Consciousness Studies* 16 (5): 9–27.

Weyhrauch, Richard W. 1980. "Prolegomena to a Theory of Mechanized Formal Reasoning." *Artificial Intelligence* 13 (1–2): 133–170.

Weyhrauch, Richard W. 1995. "Building Conscious Artifacts." In *Consciousness: Distinction and Reflection*, edited by G. Trautteur, 18–41. Napoli: Bibliopolis.

Zylberberg, Ariel, Diego Fernández Slezak, Pieter R. Roelfsema, Stanislas Dehaene, and Mariano Sigman. 2010. "The Brain's Router: A Cortical Network Model of Serial Processing in the Primate Brain." *PLoS Computational Biology* 6 (4): e1000765. https://doi.org/10.1371/journal.pcbi.1000765.