Self-awareness in Cyber-Physical Systems: Recent Developments and Open Challenges

Lukas Esterle
Aarhus University and DIGIT
Aarhus, Denmark
lukas.esterle@ece.au.dk

Nikil Dutt
University of California
Irvine, CA, USA
dutt@ics.uci.edu

Christian Gruhl
Univeristy of Kassel
Kassel, Germany
cgruhl@uni-kassel.de

Peter R. Lewis
Ontario Tech University
Toronto, Canada
peter.lewis@ontariotechu.ca

Lucio Marcenaro
University of Genoa
Genoa, Italy
lucio.marcenaro@unige.it

Carlo Regazzoni
University of Genoa
Genoa, Italy
carlo.regazzoni@unige.it

Axel Jantsch
TU Wien
Vienna, Austria
axel.jantsch@tuwien.ac.at

***Abstract*—Self-aware computing systems enable computing systems to reflect on their actions and behavior. This becomes even more relevant in Cyber-Physical Systems where computing systems have to control and interact with elements in the real world. This paper reports on recent advances made in computational self-awareness for cyber-physical systems.**

***Index Terms*—self-awareness, cyber-physical systems, reflection, autonomy**

I. Introduction

Modern computing systems are becoming more tightly integrated with their physical environment. This poses challenges as computing systems, operating in discrete time, have to deal with continuous time and space from the real world [1], [2]. At the same time, these systems are subject to unexpected changes from the environment, including other systems as well as humans [3]. While developers can make assumptions about potential changes and dynamics in the physical environment, developers will not be able to cover all possible dynamics. This leads to systems requiring an ability to sense and understand their environment. Furthermore, to ensure these systems operate to the maximum of their capacity, they require an awareness of themselves to optimize their own actions. This awareness can range from simple understanding of their algorithmic capacity to an understanding of their physical extension; from their individual actions to interactions and collaborative behaviors [4].

In this paper we summarize the contributions to the special session on Self-awareness in Cyber-Physical Systems, part of the Autonomous Design Initiative at DATE2023.

First, Peter Lewis discusses a reflective architecture for autonomous agents in Section II. Different control flows enable the agent to reflect on various aspects such as behavior, goals, or learning. Afterwards, Carlo Rigazzoni and Lucio Marcenaro propose a hierarchical framework for self-aware systems in Section III. Using Generalized Hirarchical Dynamic Baysian Networks in different cases such as creating world models, performing active inference, or generating meaningful latent variables from sensed data allows to improve self-awareness in computing systems. Third, Nikil Dutt highlights the need for a comprehensive approach to achieving adaptivity and resilience in autonomous self-aware systems in Section IV. He further argues that there are architectural components available to implement self-awareness in Cyber-physical Systems. Finally, Christian Gruhl outlines Cognitive Energy Systems (CES) in Section V. The CES represents a future version of the energy grid with higher complexity requiring a decentralized and self-aware control system. We conclude the paper with a discussion of open challenges and potential future research directions.

Identify applicable funding agency here. If none, delete this.

II. Self-Aware Machine Intelligence

The idea of self-awareness in machines has a long history, stemming mostly from science fiction, and myth before that, and often tending toward the apocalyptic. Even in research and development, the tendency to over-interpret or over-promise such mental qualities is easily done. For example, in 1958 Frank Rosenblatt, pioneer of the Perceptron artificial neural network system claimed that they could be 'conscious of their existence'[1]. More recently, Google engineer Blake Lemoine claimed that their internal AI chatbot system, LaMDA, is 'sentient'. Lemoine's statement led to a range of responses, ranging from ridicule[2] to pity[3] to concern over the growth of anthropomophism[4] to arguments that this distracts from the real ethical and power issues that currently plague AI[5]. The reactions and debate that follows these claims often causes pause for thought in the public sphere, among practitioners, and researchers. And there are valid concerns here.

Indeed, if we can learn one thing from the continual arising of such claims and the discussion or dismissal that follows, it is that we ought to be better with terminology and definitional

[1] 'New Navy Device Learns By Doing', New York Times. July 8, 1958, p25.
[2] https://twitter.com/emilymbender/status/1536198662656626688
[3] https://www.theatlantic.com/ideas/archive/2022/06/google-lamda-chatbot-sentient-ai/661322/
[4] https://www.msn.com/en-us/news/technology/timnit-gebru-and-margaret-mitchell-ai-isn-t-sentient-but-if-it-were-it-would-be-racist/ar-AAYEjaT
[5] https://www.wired.com/story/lamda-sentient-ai-bias-google-blake-lemoine/

precision. For example, self-awareness is often unfortunately conflated with consciousness or sentience in popular discussion, and it is also tempting to think that 'self-awareness' might be something mystically off-limits for engineered systems.

Yet this is not the case. Self-awareness serves a crucial function in many biological beings (including but not limited to humans), especially in uncertain, new, and social situations. Self-awareness has an evolutionary value in the world, and this accounts for its presence [5]. Fundamentally, self-awareness arises from a suite of information flows internal to an agent that adaptively feed back to help direct behavior, change goals, and set intentions [6]. Quite aside from how it might be implemented in human minds, many of the processes associated with self-awareness (and there are many) can be modeled, specified, and reproduced in machines to obtain a similar functional value [7]–[10].

To be clear, none of this requires us to accept notions such as qualia, subjective experience, or sentience in machines. It is the presence of one or more reflective 'loops' that is crucial to the function of self-awareness. The road to self-aware machine intelligence is therefore at least in-part architectural, in the creation of reflective loops and self-modeling capabilities.

In this direction, Lewis and Sarkadi [11] highlight how today's AI systems are typically not architected to contain these reflective loops, and thus do not permit self-awareness. Analyzing Critic [12] and BDI [13] agent architectures, feed-forward neural networks, and Generative Adversarial Networks (GANs) [14], they argue that none of these has the capability for reflection. Similarly, while cognitive architectures such as Practical Reasoning Systems (PRS) [15] and ACT-R [16] do not rule out reflection, they do not explicitly address it either. On the other hand, the self-adaptive and self-organizing systems research community, which has long explicitly considered self-reference (e.g. [6]), has proposed a number of high-level architectural frameworks for reflective self-awareness. One such is the LRA-M architecture due to Kounev et al. [17], while Lewis et al. [18] discuss architectural primitives that provide for a number of 'levels' of self-awareness.

Integrating LRA-M with Russell and Norvig's widely used Critic Agent architecture [12], Lewis and Sarkadi propose a 'Reflective Agent' architecture, that integrates operational learning (e.g. using reinforcement learning) with a suite of 'reflective loops'. Coupled with self-modeling capabilities, these provide for different forms of self-awareness at runtime. There are a number of self-awareness information and control flows captured in this architecture that may prove valuable in the development of cyber-physical systems operating in uncertain, dynamic, and social[6] environments. They include:

- Flow 1: Govern Behaviour. *E.g. Intervening to prevent an intended action.*
- Flow 2: Abstract Conceptualization of New Experiences. *E.g. Building new semantically rich models of itself in its environment, from experiences.*
- Flow 3: Learn about and integrate new extrinsic factors into operational goals. *E.g. Social norms, standards, and new user preferences, discovered in the environment from signs, verbal instructions, and observation of behavior.*
- Flow 4: Integrate new design goals.
- Flow 5: Active Experimentation to Improve Potential Behavior. *E.g. Proposing novel courses of action and testing hypotheses regarding them through internal simulation or in the world.*
- Flow 6: Reflecting on effectiveness of current operational goals and progress towards them. *E.g. Counterfactual reasoning about current and potential goals; asking 'am I stuck?' or 'would a different reward function better serve my high-level goals?'*
- Flow 7: Reflecting on the current mechanisms of learning. *E.g. reasoning about current operational learning mechanisms; 'could I try to learn in a different way?'*
- Flow 8: Reflective Thinking. *E.g. Refactoring and reconciling models on the fly, re-representing existing conceptual knowledge, concept synthesis.*

These essentially form a 'menu' of possible functionalities that may be afforded by this architectural approach. It may be desirable to include some, all, or none of these, depending on the form of self-awareness desired and the system's requirements and context. A given instance of a self-aware system may therefore have one or more of these processes.

In realizing this, many existing learning and reasoning algorithms may be used; the 'trick' is simply to direct the attention or input of the algorithm at an aspect of the system itself. However new techniques will also be needed. Here we highlight perhaps the largest of these: there is a need to develop mechanisms that learn human- and machine-interpretable conceptual and simulation models from empirical data and semantic information in the world, and further, for these methods to do this in an unsupervised fashion, on the fly in a complex environment. Further, to enable systems to perform internal simulations of their potential actions, goals and ways of operating, and check the likely outcomes of these before putting them into practice, there is also the need to develop the capability to run, analyze, and interpret these new models on the fly, according to need.

III. Incremental self/awareness based on free energy minimisation for autonomous agents

The definition of a computational framework allowing an autonomous agent to improve its self-awareness on the basis of its perceptual experiences while doing different tasks is a core issue of Artificial General Intelligence. Such capability can be a preliminary step to allow a more complete theory of consciousness in artificial machines to be developed, beyond being of practical use in several applications like autonomous vehicle maneuvering and cognitive wireless communications. There are few theories that provide a computational basis to consider the different aspects implied in self-awareness within

[6]We use the term 'social' here to mean the class of situations in which direct or indirect interactions exist with other agents, whether they be humans or other systems, leading to some level of common dependence, resource contention, or organization, and where intentional action directed towards these factors is necessary or valuable. See e.g., Barnes et al. [19] and Scott and Pitt [20].)

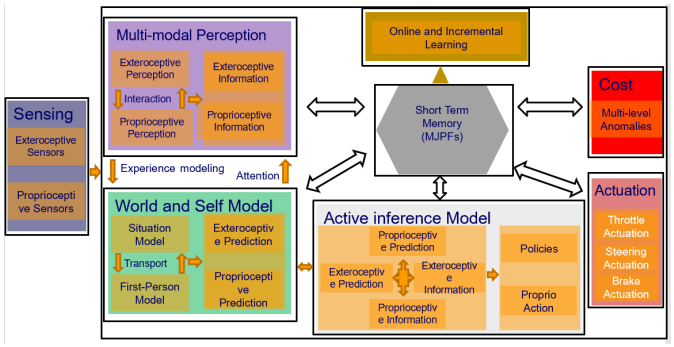

Fig. 1. Self-Awareness for Autonomous Agents

a unique framework. In [21], the main self-aware features of a self-awareness model have been listed as follows: generative modeling, discriminative modeling, interaction, hierarchical modeling, temporal reasoning, and uncertain reasoning. By integrating such properties, a self-aware agent should be able:

- to predict the not yet observed state of the world and of the self-based on its current knowledge (generative);
- to understand which of the generative models it has (or a part of it) better fits current observed data (discriminative),
- to detect if no model currently available fits current data (anomaly detection),
- to exploit knowledge of available models and to explore new ones for driving its own decision-making capabilities (support to action),
- to update by learning new models to keep its knowledge in equilibrium with changing stationary rules that drive the generation of perceptions (incremental learning)

Knowledge models for self-awareness should be capable of representing temporal variables, describing the state of the world and self at different hierarchical abstraction levels, and capturing uncertainty in representation and inference. Probabilistic graphical models, namely Dynamic Bayesian Networks (DBN), could be a good candidate for computationally defining a representation and inference basis. It has been shown [22] that a particular class of Bayesian models. i.e., Generalized Filters using generalized coordinates, allows establishing a link between probability theory and statistical mechanics that can be useful to describe a self-poietic agent capable to organize its knowledge according to a computationally defined mechanism. Generalized Hierarchical DBN (GHDBN) has been proposed in [21] as a core computation tool for self-aware model definition. In particular, free energy minimization can be used to explain how an agent can do all steps necessary to make itself aware, guided by the principle of keeping the homeostatic equilibrium with data generated by the world and itself. The driving principle from statistical mechanics is the least action principle: to minimize the work to be done, a filter tries to keep constant the action included in a model where rules remain stationary.

In the framework described in Fig. 1 one can see how the different parts of the self-awareness model should include related capabilities at a macro level.

- Multi-modal Perception (GHDBNs allowing the sensing data of the agent to be converted into meaningful latent variables)
- World and Self models (autobiographical memories in lower dimensionality GHDBNs)
- Active inference Model (GHDBNs integrated with action variables useful to drive decision-making and actuators)
- Cost block (including variables useful to evaluate free energy using current models on current data and to compute measurements useful to update model themselves.
- Online and Incremental Learning (machine learning methods to produce new GHDBNs solving free energy minimization)
- Short-term memory (as the dynamic bayesian process realizing inference oh GHDBns so producing new knowledge like generative, discriminative estimates using bayesian inference, anomaly detection, and make them available to appropriate modules)

Self-awareness frameworks, as suggested here, should be part of the research in the coming years to allow autonomous agents to improve their capability of interacting with humans in terms of explainability, adaptation, and coordination.

IV. Adaptive, Resilient Computing Platforms through Self-Awareness

At its very core, a CPS deploys a primitive ODA loop for its operation: *Observe*: sense environmental data in the physical world; *Decide*: use computing platforms for sense-making; and *Act*: deploy actuators in the physical world. This primitive loop is akin to a basic feedback control loop, where decisions are made primarily based on a history of past observations and actions. This approach may have worked well for relatively simple use cases with known expected behaviors, static operational modes, and predictable (or expected) environmental scenarios. However, a contemporary CPS faces an explosion in diversity and complexity across the entire computing stack, from low-level hardware/computing architectures, to the highest level of applications and policies, that pose tremendous challenges for supporting adaptivity and resilience. Indeed, today's CPS are stitched together – often in an ad-hoc manner – using an array of algorithms, policies, and (increasingly) black-box machine learning models in an effort to deal with increasing dynamism, unmodeled/anomalous behaviors, and critical system failures. The pace of innovations in computing hardware and software has exacerbated this problem, given the explosion of diversity and complexity in emerging CPS across the abstraction stack:

- At the highest level of computing abstraction, applications exhibit complex dynamic behaviors that span a diverse scale of complexity from small footprint/edge/IoT devices to large systems-of-systems such as interacting groups of autonomous systems (e.g., drone swarms and truck platoons) to systems-at-scale (large data centers). These CPS must handle dynamic behaviors and yet-unseen behaviors in use-cases, contexts, sensed data, as well in actuation. In turn, the sensed data as well as the processed outputs are truly heterogeneous in type (static, periodic, streams, events) as well as in criticality requirements.

- The lowest computing abstraction levels of technology and architecture are evolving rapidly, driven by the need for customization to meet the possibly conflicting design goals of performance, energy, resilience, etc.; as well as the emergence of newer device and memory technologies. The landscape of processor architectures is changing rapidly, moving from standard CPUs, to GPUs, reconfigurable hardware, and domain-specific accelerators. Memory and interconnect are evolving rapidly as well, with the mainstream acceptance of Non-Volatile Memories (NVM) and newer storage technologies, spanning the gamut from local to cloud storage, and evolving interconnect fabrics.
- Software toolchains: the diversity in computing platforms results in a diverse and tessellated software ecosystem across virtualization schemes, runtime systems, compilers, code generators, etc.
- Emergent behaviors arising from collections of multiple interacting (and increasingly autonomous) CPS, resulting in a cross-product of individual known and unknown behaviors that at best are impossible to specify at design time, and at worst manifest as emergent, yet-unseen behaviors that may not only compromise mission safety, but which may initiate a chain of system-wide critical failures that jeopardize other CPS entities.

Computational self-awareness principles show promise for achieving adaptivity and resilience in the face of such diversity and dynamism. The rich history of deploying self-X principles in computing platforms is captured in this incomplete list:

- The testing and fault-tolerant computing community has studied self-test, self-diagnosis, self-repair for chips and computing platforms for several decades; and unmanned space missions have embraced some self-X principles to ensure high levels of mission resilience.
- The software community leveraged concepts of reflective software in the early 2000's and IBM led a highly visible program in Autonomic Computing (IBM) [23].
- In the past decade, the computer architecture community has studied ODA-based feedback control strategies for software centric and targeted adaptation of computing platform resources [24], [25].

While these efforts have contributed to the development of primitive self-aware computational platforms, they lack a comprehensive approach to achieving adaptivity and resilience through a principled treatment of computational self-awareness that typically cover:

- Cross-layer sensing and actuation [26].
- Self-models and environment models that experience phenomena and are aware of state and behaviors [27].
- Introspection that combines both reactive and proactive/reflective control loops [28]. Reactive loops operate in a classical ODA control loop, based purely on past behaviors. Proactive/Reflective loops enable proactive control behaviors, that consider both past as well as possible future outcomes using an Observe-Reflect-Decide-Act (ORDA) loop. The ODA and ORDA loops operate concurrently to enable effective introspection for complex systems, similar to the fast, autonomic nervous system in animals (ODA) and the slower reflective system (ORDA) that enables planning, policies, strategies and evaluation of alternatives.
- A hierarchy of ODA and ORDA loops at different time scales in a CPS [29], that model the entire spectrum of computing abstraction levels from high-level applications, through software, and down to the architectural hardware.
- Adaptive behaviors driven by models of external and internal environment. This requires explicit modeling of goals, constraints and evolvable policies to effect actions in the CPS [30].

This is a rich and active research area with researchers proposing models, architectures, and design flows for computational self-awareness, special journal issues [31]–[33] and many workshop series such as SelPhyS [34] that attempt to understand and harness computational self-awareness principles and their applications across different domains. For instance, we have used the metaphor of self-aware information processing factories (IPF) [35] to operate future adaptive computational platforms akin to sensor-actuator-rich factories, exploiting online optimization using self-reflection and self-organization to enable system autonomy in the face of dynamic changes in workload and operational environments. IPF principles aim for flexible, adaptive management of computational resources while ensuring continuous operation, achieved through maximally distributed autonomy with minimized centralized control.

Another perspective is embodied self-aware computing [36]. Since computational platforms for CPS are embodied in the physical environment, we build on the notion of embodied cognition to describe embodied self-aware computing systems, where the computing platform can – similar to a brain embodied in the environment – operate as an agent in a physical world and achieve complex and dynamically changing goals. This can be achieved through architectural components that facilitate: learning, reasoning, and managing complex, dynamically changing goals. Self-aware CPS systems can then be engineered using template reference architectures that deploy the following embodied self-aware architectural components:

- Models for introspection, covering both self-models as well as external environmental models; these models need to capture history, and must have the ability to adapt dynamically through learning mechanisms
- Explicit modeling of goal hierarchies and control mechanisms to achieve adaptive actions, and possibly conflicting goals.
- Assessment and attention mechanisms to enable calibration of internal and external states, as well as ability to react to major external changes or anomalous behaviors.
- Learning mechanisms across all architectural components: introspective models, goal hierarchies, control, assessment and attention; enabling specialization, customization and system resilience.

While there is a rich history of self-aware computing efforts, CPS designers must use these principles judiciously, since they exact overheads in resources, performance and energy. Furthermore, the use of black-box machine learning techniques

may compromise resilience due to the lack of explainability, or worse yet, through malicious exploitation of these techniques. We posit that to fully exploit computational self-awareness for achieving adaptive, resilient computing platforms, we need principled approaches to flexibly incorporate the following concepts: Cross-layer sensing/actuation; Self- and Environmental models that learn and evolve over time in an explainable manner; and Control/coordination of emergent behaviors for safe and possibly predictable outcomes to achieve dynamic goals as well as system resilience. Efforts in run-time verification may be useful to complement other formal efforts for static and dynamic system analysis.

V. Cognitive Energy Systems

What are Cognitive Energy Systems (CES)? A cognitive system describes a system that has an awareness of its possibilities of action, can perceive its environment and can independently work out solutions for changing tasks through analysis, learning and problem-solving mechanisms. This definition corresponds, in essence, to our understanding of self-awareness. As of now, our power grid is primarily a centrally controlled system. However, this will change to a decentralized system with many heterogeneous actors. One reason for this is the increasing number of small "power plants" (e.g. PV plants or wind power plants) that contribute to energy production but simultaneously belong to different stakeholders. Consumers are also changing, for example, electric vehicles. They have a high charging consumption but, at the same time, can be used as energy storage. As a result, our power grid will become increasingly complex and, at the same time, more flexible. CES is one possible answer to making this complexity manageable again, but it comes with its own open challenges.

In [37], we present a vision of tackling the aforementioned complexity using the Organic Distribution System (ODiS). Here, organic refers to *organic-computing* [25]. It introduces a hierarchy consisting of Organic Home Energy Management Systems (O-HEMS) operated at the customer's site (e.g. the end user at home). A large number of *prosumers* (e.g. electric vehicles) are coordinated by these HEMS to ensure that they are operating within the operational limits of the (low voltage) grid. The low voltage grid is managed by an Organic Distributed Management System (O-DMS), which aims to keep the power system in a normal operational state. All agents are connected to a computer network, allowing the exchange of information. Each system (i.e. O-HEMS and O-DMS) can interact with its environment, e.g. by charging electric vehicles, adjusting dynamic loads, or injecting energy into the grid (O-HEMS) or by changing the grid topology by controlling switches and substations (O-DMS). This flexibility comes with an increased complexity since the decisions of each agent influence the environment of other agents. The idea of self-awareness is a promising approach to managing the increased complexity. Here we understand self-awareness as the ability to investigate the system's own condition (see also self-reflection [38]) and the condition of its environment and to be able to make decisions and assess their effects if unexpected or sudden changes are detected.

Research on self-awareness for CES has only just begun. Simulation-based solutions for decision-making will undoubtedly play a significant role. This is already becoming apparent in challenges such as L2RPN [39], in which a power grid is to be controlled autonomously (the challenge is becoming more difficult every year). In addition, further research into novelty detection techniques will play an essential role as it can be seen as one of the fundamental building blocks of awareness in technical systems [40]. The distributed nature of CES' and the mutual influence of the participating agents makes CES' an application domain for self-integration research [41]. Another critical component to making systems *self-aware* is that reasonable computing resources are available. It is foreseeable that a home controller (such as the previously presented O-HEMS) will have significantly less computing power than, for instance, a large wind park. The challenge here is to make decisions even on such *microsystems* and implement at least (partial) self-awareness.

Self-awareness in CES' will be a crucial building block to increase resilience against failures and attacks. The recent past, unfortunately, shows how vital resilience is for critical infrastructure and that targeted attacks could become more and more likely. In addition to direct attacks, such as on the Nord Stream pipelines[7], communication networks can also be targeted, such as the sabotage against the German railway's GSM-R network in October 2022[8].

Here, we only briefly overview CES' and its related research fields. It should be clear that CES' will be one of the CPS' that would highly benefit from future self-awareness techniques.

VI. Conclusion

Self-awareness in Cyber-Physical Systems is a rich research area with contributions from different fields. In addition to the already mentioned challenges, we can further identify a set of challenges necessary to address in research:

- **Runtime modeling and model calibration:** Models are essential for self-awareness. While we have extensive models for physical properties they are often imprecise. Utilizing model calibration techniques allow for improving initial models [42]. With the increased interest in digital twins, continuous time models and accompanying calibration approaches have seen a push recently [43].
- **Mutual modeling and causality:** For reasoning about and predicting the behavior of other agents in its environment, a self-aware system needs to build and maintain models of those agents [4], [44].
- **Verification and trust:** To improve trust in autonomous and self-aware CPS, systems are required to be verifiable [45]. Verification monitors, however, need to be defined before deployment. Nevertheless, if we can verify behavior and guarantee actions and outcomes, this will also lead to increased trust from users.
- **Distributed self-awareness:** With the rise of the Internet of Things and its spread into various industries, swarm

[7]https://www.dw.com/en/a-63806519 (last access 05.12.2022)
[8]https://www.dw.com/en/a-63377385 (last access 05.12.2022)

behavior and self-organization received a push. Allowing systems to share their knowledge and models across multiple devices may allow them to generate those models faster and more accurately. A question remains whether we can achieve this also with computational self-awareness.

REFERENCES

[1] E. A. Lee, "Cyber physical systems: Design challenges," in *Proc. of the Int. Symp. on object and component-oriented real-time distributed computing*. IEEE, 2008, pp. 363–369.

[2] L. Esterle and R. Grosu, "Cyber-physical systems: challenge of the 21st century," *e&i*, vol. 133, no. 7, pp. 299–303, 2016.

[3] K. L. Bellman, C. Landauer, N. D. Dutt, L. Esterle, A. Herkersdorf, A. Jantsch, N. Taherinejad, P. R. Lewis, M. Platzner, and K. Tammemäe, "Self-aware cyber-physical systems," *ACM Trans. Cyber Phys. Syst.*, vol. 4, no. 4, pp. 38:1–38:26, 2020.

[4] L. Esterle and J. N. A. Brown, "I think therefore you are: Models for interaction in collectives of self-aware cyber-physical systems," *ACM Trans. Cyber Phys. Syst.*, vol. 4, no. 4, pp. 39:1–39:25, 2020.

[5] C. A. Lage, D. W. Wolmarans, and D. C. Mograbi, "An evolutionary view of self-awareness," *Behavioural Processes*, vol. 194, p. 104543, 2022.

[6] P. R. Lewis, A. Chandra, S. Parsons, E. Robinson, K. Glette, R. Bahsoon, J. Torresen, and X. Yao, "A Survey of Self-Awareness and Its Application in Computing Systems," in *Proc. of the Int. Conf. on Self-Adaptive and Self-Organizing Systems Workshops*. IEEE Computer Society, 2011, pp. 102–107.

[7] P. R. Lewis, M. Platzner, B. Rinner, J. Torresen, and X. Yao, *Self-Aware Computing Systems: An Engineering Approach*, P. R. Lewis, M. Platzner, B. Rinner, J. Tørresen, and X. Yao, Eds. Springer, 2016.

[8] P. R. Lewis, "Self-aware computing systems: From psychology to engineering," in *Proc. of the Design, Automation & Test in Europe Conference & Exhibition*, 2017, pp. 1044–1049.

[9] S. Kounev, J. O. Kephart, A. Milenkoski, and X. Zhu, Eds., *Self-Aware Computing Systems*. Springer, 2017.

[10] C. Blum, A. F. T. Winfield, and V. V. Hafner, "Simulation-based internal models for safer robots," *Frontiers in Robotics and AI*, 2018.

[11] P. R. Lewis and S. Sarkadi, "Reflective artificial intelligence," 2023. [Online]. Available: https://arxiv.org/abs/2301.10823

[12] S. Russell and P. Norvig, "Artificial intelligence: A modern approach," *Foundations*, vol. 19, p. 23, 2021.

[13] A. S. Rao, M. P. Georgeff *et al.*, "BDI agents: From theory to practice," in *ICMAS*, vol. 95, 1995, pp. 312–319.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[15] B. C. Smith, "Reflection and semantics in lisp," in *Proc. of the Symposium on Principles of programming languages*, 1984, pp. 23–35.

[16] J. R. Anderson, M. Matessa, and C. Lebiere, "Act-r: A theory of higher level cognition and its relation to visual attention," *Human–Computer Interaction*, vol. 12, no. 4, pp. 439–462, 1997.

[17] S. Kounev, P. Lewis, K. Bellman, N. Bencomo, J. Camara, A. Diaconescu, L. Esterle, K. Geihs, H. Giese, S. Göetz, P. Inverardi, J. Kephart, and A. Zisman, "The notion of self-aware computing," in *Self-Aware Computing Systems*, S. Kounev, J. O. Kephart, A. Milenkoski, and X. Zhu, Eds. Springer, 2017, pp. 3–16.

[18] P. R. Lewis, A. Chandra, F. Faniyi, K. Glette, T. Chen, R. Bahsoon, J. Torresen, and X. Yao, "Architectural aspects of self-aware and self-expressive computing systems," *IEEE Computer*, vol. 48, pp. 62–70, 2015.

[19] C. M. Barnes, A. Ekárt, and P. R. Lewis, "Social action in socially situated agents," in *Proc. of the Int. Conf. on Self-Adaptive and Self-Organizing Systems*, 2019, pp. 97–106.

[20] M. Scott and J. Pitt, "Inter-dependent self-organising mechanisms for cooperative survival," 2023, in Press.

[21] C. S. Regazzoni, L. Marcenaro, D. Campo, and B. Rinner, "Multisensorial generative and descriptive self-awareness models for autonomous systems," *Proc. of the IEEE*, vol. 108, no. 7, pp. 987–1010, 2020.

[22] K. Friston, B. Sengupta, and G. Auletta, "Cognitive dynamics: From attractors to active inference," *Proc. of the IEEE*, vol. 102, no. 4, pp. 427–445, 2014.

[23] J. Kephart and D. Chess, "The vision of autonomic computing," *Computer*, vol. 36, no. 1, pp. 41–50, 2003.

[24] H. Hoffmann, M. Maggio, M. D. Santambrogio, A. Leva, and A. Agarwal, "Seec: A framework for self-aware computing," MIT, Cambrige, Massachusetts, Tech. Rep. MIT-CSAIL-TR-2010-049, October 2010.

[25] C. Müller-Schloer and S. Tomforde, *Organic Computing-Technical Systems for Survival in the Real World*. Springer, 2017.

[26] S. Sarma, N. Dutt, P. Gupta, N. Venkatasubramanian, and A. Nicolau, "Cyberphysical-system-on-chip (cpsoc): A self-aware mpsoc paradigm with cross-layer virtual sensing and actuation," in *Proc. of the Design, Automation & Test in Europe Conference & Exhibition*, 2015, pp. 625–628.

[27] A. Jantsch, N. Dutt, and A. M. Rahmani, "Self-awareness in systems on chip— a survey," *IEEE Design & Test*, vol. 34, no. 6, pp. 8–26, 2017.

[28] N. Dutt, A. Jantsch, and S. Sarma, "Toward smart embedded systems: A self-aware system-on-chip (soc) perspective," *ACM Trans. Embed. Comput. Syst.*, vol. 15, no. 2, feb 2016.

[29] A. M. Rahmani, A. Jantsch, and N. Dutt, "Hdgm: Hierarchical dynamic goal management for many-core resource allocation," *IEEE Embedded Systems Letters*, vol. 10, no. 3, pp. 61–64, 2018.

[30] A. Jantsch, A. Anzanpour, H. Kholerdi, I. Azimi, L. C. Siafara, A. M. Rahmani, N. TaheriNejad, P. Liljeberg, and N. Dutt, "Hierarchical dynamic goal management for iot systems," in *Proc. of the Int. Symp. on Quality Electronic Design*, 2018, pp. 370–375.

[31] A. Jantsch, P. R. Lewis, and N. Dutt, "Introduction to the special issue on self-aware cyber-physical systems," *ACM Trans. Cyber-Phys. Syst.*, vol. 4, no. 4, jun 2020.

[32] J. Torresen, C. Plessl, and X. Yao, "Self-aware and self-expressive systems," *Computer*, vol. 48, no. 07, pp. 18–20, jul 2015.

[33] N. Dutt, C. S. Regazzoni, B. Rinner, and X. Yao, "Self-awareness for autonomous systems," *Proc. of the IEEE*, vol. 108, no. 7, pp. 971–975, 2020.

[34] "SelPhyS: Self-Awareness in Cyber-Physical Systems," https://selphys.ict.tuwien.ac.at/.

[35] N. Dutt, F. J. Kurdahi, R. Ernst, and A. Herkersdorf, "Conquering mpsoc complexity with principles of a self-aware information processing factory," in *Proc. of the Int. Conf. on Hardware/Software Codesign and System Synthesis*, 2016.

[36] H. Hoffmann, A. Jantsch, and N. D. Dutt, "Embodied self-aware computing systems," *Proc. of the IEEE*, vol. 108, no. 7, pp. 1027–1046, 2020.

[37] I. Loeser, M. Braun, C. Gruhl, J.-H. Menke, B. Sick, and S. Tomforde, "The vision of self-management in cognitive organic power distribution systems," *Energies*, vol. 15, no. 3, 2022.

[38] S. Tomforde, J. Hähner, S. von Mammen, C. Gruhl, B. Sick, and K. Geihs, ""know thyself" - computational self-reflection in intelligent technical systems," in *Proc. of the Int. Conf. on Self-Adaptive and Self-Organizing Systems Workshops*, 2014, pp. 150–159.

[39] A. Marot, B. Donnot, G. Dulac-Arnold, A. Kelly, A. O'Sullivan, J. Viebahn, M. Awad, I. Guyon, P. Panciatici, and C. Romero, "Learning to run a power network challenge: a retrospective analysis," in *Proc. of the NeurIPS 2020 Competition and Demonstration Track*, ser. Proc. of Machine Learning Research, H. J. Escalante and K. Hofmann, Eds., vol. 133. PMLR, 06–12 Dec 2021, pp. 112–132.

[40] C. Gruhl, B. Sick, A. Wacker, S. Tomforde, and J. Hähner, "A building block for awareness in technical systems: Online novelty detection and reaction with an application in intrusion detection," in *IEEE iCAST*, 2015, pp. 194–200.

[41] K. Bellman, J. Botev, A. Diaconescu, L. Esterle, C. Gruhl, C. Landauer, P. R. Lewis, P. R. Nelson, E. Pournaras, A. Stein, and S. Tomforde, "Self-improving system integration: Mastering continuous change," *Future Generation Computer Systems*, vol. 117, pp. 29–46, 2021.

[42] C. Semeraro, M. Lezoche, H. Panetto, and M. Dassisti, "Digital twin paradigm: A systematic literature review," *Computers in Industry*, vol. 130, p. 103469, 2021.

[43] E. Madsen, D. Tola, C. Hansen, C. Gomes, and P. G. Larsen, "Aurt: A tool for dynamics calibration of robot manipulators," in *Proc. of the Int. Symp. on System Integration*, 2022, pp. 190–195.

[44] L. Esterle, C. Gomes, M. Frasheri, H. Ejersbo, S. Tomforde, and P. G. Larsen, "Digital twins for collaboration and self-integration," in *Proc. of the Int. Conf. on Autonomic Computing and Self-Organizing Systems Companion*, 2021, pp. 172–177.

[45] L. A. Dennis and M. Fisher, "Verifiable self-aware agent-based autonomous systems," *Proc. of the IEEE*, vol. 108, no. 7, pp. 1011–1026, 2020.