Queryable Self-Deliberating Dynamic Systems

Giuseppe De Giacomo
Università degli Studi di Roma “La Sapienza”
DIAG, Via Ariosto 25, 00185, Roma, Italy
degiacomo@diag.uniroma1.it

We are witnessing an increasing availability of **autonomous systems** that operate in nondeterministic (uncertain) environments and offer some form of programmability. These include manufacturing devices, smart objects and spaces, intelligent robots, dynamic business process management systems, and many others. All these autonomous systems are currently being revolutionized by advancements in sensing (vision, language understanding) and actuation components (automated mobile manipulators, automated storage and retrieval systems). However, such autonomous systems are held back by the fact that their logic is still based on hard-wired rules either designed or possibly obtained through a learning process. On the other hand, we can envision systems that are able to **program themselves**, automatically tailor their behavior so as to achieve desired goals, maintain themselves within safe boundaries in a changing environment, and follow regulations and conventions that evolve over time. Crucially, empowering autonomous systems with self-programming carries significant risks and therefore we must be able to **balance power with safety**. For this reason we need to realize autonomous systems that are **white-box**, that is, whose behavior is at any moment fully analyzable and comprehensible in human terms, and guarded by human oversight.

To make this ideas more concrete consider the following scenario.

> *After a long week-end, the human supervisor inspects the manufacturing system and notices that a production line has slowed down significantly, though it is still producing. She queries the system on what it is doing. The system reveals to her the revised process, which is avoiding the use of the production island 176-176 by repurposing the tools in island 176-671 and sending items there. She then queries why the system has reprogrammed itself to do so. The system answers by showing that on Sunday 11:43pm the island 176-176 started to produce an unacceptable percentage of defective items, based on tests performed during production. So, the system restructured the process to achieve the specified objectives (quality and throughput) as it could in the presence of the faulty island, instead of shutting down the production line: the system analyzed the available capabilities and reprogrammed itself resulting in the current revised process.*

In the scenario above we have a **autonomous system** (the manufacturing system) with multiple components, some of which possibly use Machine Learning (ML), to provide sensing and acting capabilities (detecting defective items, reconfiguring tools parameters), that can be suitably organized to enact a *dynamic behaviour* (the production process) to meet its *specifications* (constraints on the product and production model). The system has the ability to monitor and detect faulty parts of the process, by acquiring semantically rich **first-order/relational data** about objects of interest, their properties and their relationships. Crucially, the system has **self-programming** abilities that it can use to modify its current behavior, i.e., its logic. Notably, the system can be *queried* to analyze, in terms understandable to humans, under which circumstances and for fulfilling which *specifications*, it reprogrammed itself. Moreover it can be queried on whether its *self-synthesized program* meets any dynamic property of interest to the human supervisor (e.g., for “what-if” analysis). In a slogan, the system is **white-box**. This vision can become a reality. To do so *we need to lay the theoretical foundations and developing practical methodologies of a science and engineering of* ***white-box self-programming autonomous systems***. Notice that this program is also in line with the recent vision that DARPA wants to address within the $2B “AI next” initiative: “Today, machines lack contextual reasoning capabilities, and their training must cover every eventuality, which is not only costly, but ultimately impossible. We want to explore how machines can acquire human-like communication and reasoning capabilities, with the ability to recognize new situations and environments and adapt to them.”[1]

The significance of this enterprise can be understood within three areas of pivotal importance in the current socio-economic context, namely:

1. **Smart Manufacturing**, where significant research efforts are focusing on improving flexibility, agility and productivity of manufacturing systems, under the umbrella term *Industry 4.0*, or *4th industrial revolution*.[2]
2. **Internet of Things**, which is rising as a virtual fabric that connects “things” equipped with chips, sen-

[1] https://www.darpa.mil/news-events/2018-09-07.

[2] M. Lorenz et al. *Man and Machine in Industry 4.0: How Will Technology Transform the Industrial Workforce Through 2025?* The Boston Consulting Group. 2015.

sors and actuators and allows for building **smart objects** and **smart spaces** with high levels of awareness of the environment and its human occupants.[3]

3. **Business Process Management**, which advocates explicit conceptual descriptions of a process to be enacted within an organization or possibly across organizations, and which is instrumental to business processes improvement, the top business strategy of CIOs according to Gartner. [4]

We observe that forms of self-programmability have been advocated in all the above contexts. For example, it is advocated that cyber-physical systems in Manufacturing or Internet of Things should be able to **adapt** their logic to deal with changes in their environment by **exploiting information gathered at runtime**. However, it is considered **impossible to determine a priori all possible adaptations** that may be needed; thus self-programming abilities would be highly desirable. In Business Processes, it is considered important for the next generation of process management systems to allow processes to automatically **recover** when unanticipated exceptions occur, without explicitly defining a priori **recovery policies**, and **without the intervention of domain experts** at runtime. These self-programming abilities would reduce costly and error-prone manual ad-hoc changes, and would relieve software engineers from mundane adaptation tasks. Note that some of these concerns have been shared by *autonomic computing*, which has promoted self-configuration, self-healing, self-optimization, and self-protection, though by using policies provided by IT professionals. Sophisticated languages and methodologies for streamlining the development of adaptation and exception handling have been devised, however IT professionals still need to write the logic of these through **hard-wired rules encoded in hand-crafted programs**. Although interest is evident, currently self-programming abilities are missing or very limited in actual systems operating in nondeterministic, uncertain, unpredictable environments.

I believe that, by exploiting advances in AI, and in particular Knowledge Representation, Planning and Synthesis, Agent Technologies, Automated Reasoning, Reinforcement Learning and forms of Deep Learning, in the future years it will be possible to **equip autonomous systems with advanced self-programming abilities**. So as to enable systems to:

- **Achieve desired goals**, that is guarantee that a certain desired state of affairs is eventually reached. In the above example a manufacturing system automatically reconfigures the fabrication process if some workstation is producing too many defective items, by changing the sequencing of processing units so as to temporarily cutout the defective tool from the process, thus achieving an acceptable error rate.
- **Maintain themselves within a safe boundary** against unanticipated changes in the environment in which they operate. For example the system of a smart building may keep the desired temperature and humidity in a museum room at some desired level, even in presence of an unforeseen large crowd of visitors, possibly by momentarily repurposing other actuators, such as a secondary air conditioning system typically used only in case of failure of the main one.
- **Keep following regulations and conventions** that evolve over time while enacting their behavior. For example, to satisfy a new privacy regulation, a business process may refine its behavior to guarantee that the sensitive data are erased from the system before the completion of each process instance.

More generally, we intend to **enable autonomous systems to act in an informed and intelligent way in their environment**, by changing the way they behave as a consequence of the information they acquire from the external world and exchange with the humans operating therein.

Since "*with great power comes great responsibility*", introducing advanced forms of self-programming calls for the ability to make the behavior automatically synthesized by the autonomous system **comprehensible** to human supervisors, who are thus able to control and guide it. So it is indeed crucial to develop self-programming autonomous systems that **can be queried**, i.e., that are **white-box**: in every moment the system can be queried for the status of its specifications, and on whether its behavior meets any dynamic property of interest to the human supervisors. Ultimately it is the fact that the resulting behavior is **analyzable in terms comprehensible to humans** that will make white-box self-programming autonomous systems **trustworthy**.

We stress that the need to move towards **white-box** approaches is advocated by a large part of the **AI community**[5] as well as the **CS community**[6], and has been recently taken up by Europe, as testified by the ""Declaration, Cooperation on Artificial Intelligence", Brussels, April 10, 2018[7], by "Artificial Intelligence for Europe: Communication From The Commission To The European Parliament", COM(2018) 237, SWD(2018) 137, the European Commission, April 25, 2018[8], and by the CLAIRE initiative of the European AI scientific community[9].

I intend to bring about this vision at Ital-AI, putting forward industrial opportunities as well as scientific challenges.

[3] C. MacGillivray et al. *Worldwide Internet of Things Forecast Update, 2016-2020*. IDC. Doc # US40755516. 2016.

[4] Gartner Group. *BPM Survey Insights*. Gartner Report. http://www.gartner.com/it/page.jsp?id=1740414.

[5] S. Russell, D. Dewey, and M. Tegmark. Research priorities for robust and beneficial artificial intelligence. AI Magazine, 36(4), 2015.

[6] ACM U.S. Public Policy Council and ACM Europe Policy Committee. Statement on algorithmic transparency and accountability. ACM, 2017.

[7] http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=50951

[8] http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=51625

[9] https://claire-ai.org