TUM

TECHNISCHE UNIVERSITÄT MÜNCHEN
INSTITUT FÜR INFORMATIK

The Dawn of the Age of Autonomy

Alois Knoll

TUM-I144

"If humanity can just get past the next 200 years without driving itself to extinction, then we're good to go."
Stephen Hawking, 2010

The Dawn of the Age of Autonomy

Alois Knoll

The 20th century certainly has seen technological changes of epic proportions, many of which were previously inconceivable. So – what will the future bring? To a much higher degree than currently possible, the future will see devices, appliances and hierarchies of complex cyber-physical systems that can govern and control their own behaviour – up to a level, where they can independently make informed decisions whose consequences may directly affect the lives of humans. This highly disruptive technology holds enormous potential, and may even be an indispensable toolset for the future survival of humanity. At the same time, because it will potentially have a deep impact on our own autonomy as human beings, technological solutions will have to be developed in parallel which ensure that individual humans and human society as a whole is able to maintain its freedom to decide. The European Union, with its diverse cultural heritage and wide-spread technology skills, is in a superb position to take the lead in this next wave of all-encompassing technology – which will have more impact on our lives than digitalization and networking has had in the last 50 years. However, in order to be in the pole position for this and become the trendsetter again, Europe needs to make the right investments now – in such a way that intelligently capitalizes on the investments in base technologies that have already been made (hardware/software/commodity components) and that concentrates on the highest level of realizing autonomy functions. This effort should be accompanied by standardization efforts that help in the market development of products using basic and increasingly advanced functions – with GSM and MPEG being prominent past examples.

Autonomy: what is it – and what is it good for?

There has been an inflationary – and often misleading – use of the term "autonomy", much like with "intelligence", another term often misused in conjunction with technology. While intelligence is a fuzzy term with many possible definitions that make it easy to tweak to one's needs, the definition of autonomy is less controversial. In the context of humans and societies, it is the capacity "of a rational individual to make an informed, un-coerced decision" and/or to "give oneself his own rules" according to which one acts in a constantly changing environment. It also means the "self-government of the people".

Technical systems that claim to be autonomous will be able to perform the necessary analogies of "mind functions" that enable humans to be autonomous, including perception of the environment, reasoning, planning, modelling, memory, and many more. Moreover, like in human and animal societies, they will have to abide by rule sets (or "law systems") that govern their interaction – between individual members and between groups.

When transferring the highly sophisticated and multi-faceted concept of autonomy to technical systems, a lot can be learned from the evolution of biological models. Not surprisingly, when looking for a way to have networked computers manage themselves without direct human intervention, in 2001,

IBM took inspiration from the human autonomic nervous system (which is part of the human peripheral nervous system), which controls visceral functions in the human body (typically without brain intervention and far below the level of consciousness). Going from there to arriving at the level of the human cognitive system, and then going even further to achieve full "decision power" (combined with access to an amount of information far beyond what any human is able to grasp) – and all that in real time – will clearly be a huge step. However, due to the progress in robotics, computational neuroscience and high performance computing, this vision is about to become reality – and sooner than we think!

There are many examples of autonomous systems that can be envisaged (which may even, by the way, help maintain our European lifestyle in the face of adverse demographic change, energy shortages, etc.), just to name a few:

- **Zero accident** passenger cars **without steering wheels**
- Decentralized, **resilient, self-healing, smart energy grids** with **zero blackout** (recovery) **time**
- **Autarkic "city brains"** with the ability to control all material/communication flows within a city – optimized according to various target functions

The advent of technical systems that exhibit true autonomy of their own will, nevertheless, not come overnight nor will it come effortlessly. As always in technology, there will be several phases in the evolution; in fact, they will be called "disruptive" when we look back at their introduction years from now. And clearly, massive investments will be needed to take the lead!

Evolution of Autonomy: what are potential steps?

From today's point of view, the following steps, which build on each other, are likely to be taken in the years to come to create increasingly autonomous systems – synchronal with and in response to growing human needs:

a) **Perception-based systems**: systems with sensors of different kinds which can pick up and combine environmental information, so as to respond to a user's needs who shares a (temporary) environment with that device. Most mobile devices in use today are capable of sensing at least where they are – and they can draw additional information that may be relevant to the user. Such information may be map data, additional information about local buildings, etc. The fusion of the information from different sources, in different data representations (and from different physical domains) is a topic of active research – and will certainly remain so for the foreseeable future. The challenge here is to *produce value-added information that precisely fills the need of the user* – and does not leave him alone in a sea of millions of web-pages.

b) **Systems with context-awareness and context interpretation**: the device will know (and/or anticipate) in which situational context it is (e.g., in a car that is slowing down or in the purse of a user who is walking in a shop) and what its owner is doing – or is going to do. It also knows in what state (of mind) the user is, and it can reason about his intentions. It also has a certain sense of "self", i.e., it knows about its own information processing and presentation power and its potential consequences, as well as the relations between itself and the environment. It may memorize consequences from past events, it may even abstract from them and transfer this gained knowledge to similar situations in the future. Combining all of this, such systems may not only present information to the user, but beyond this, they can also *analyse* and *interpret it* in a *given situational context.* They can produce *individualized knowledge* and give *specific feedback* to their interaction partners. This can be considered as *networked context-aware, real-time information analysis, interpretation and synthesis*. Simple instances of such devices are mobile phones or glasses/goggles that observe their owner. More complex exam-

ples are cars that analyse the behaviour of their passengers and combine the results with observations of the environment, as well as with models of the terrain, physics and knowledge drawn from the web in real-time. The systems are still, however, restricted to perception, reasoning, and (multimodal) information output production.

c) **Perception → Cognition → Action (PCA)-systems**: the next stage of the evolution are systems that can act in the real world, i.e., move their body based on their own planning and/or manipulate objects. This is a completely new quality of technical systems: humans allow them the partial or full freedom to decide about physical changes in our man-made, real world. By *acting in the real world*, these systems can physically change the(ir) environment, they can also gather information by systematically and deliberately moving their sensors into new positions, and *they can learn*. This implies that such artefacts need not only perceive and observe the world, but that they *make decisions at many levels of cognition*: from deciding between possibly conflicting goals about what path should be pursued, to deciding what tool to use and which way to go. *They will eventually need to simulate all the high-level intelligence functions of humans*. A somewhat "reduced" form is *partial autonomy,* where the decisions made by the systems are counter-checked by humans – over shorter or longer time spans. An important sub-class are *joint-action systems*, which work directly with humans in physical cooperation. Since these systems act directly (and not only indirectly through the human as a medium), putting them in a situation where they potentially could do harm to the world, we clearly need a set of generally applicable rules and laws that govern these systems.

d) **Societies of PCA-systems**: once we master the technology of the PCA as an individual artefact, the next (and final) phase will consist of interworking, tightly cooperating "societies" of PCA systems – all with direct access to the whole of cyberspace. The societal models for such sets of systems can be simple systems like the popular example of ant colonies that exhibit some kind of "swarm intelligence", or any other suitable heterarchical or hierarchical form of society from human or animal history. Clearly, while such societies can perform very complicated and complex tasks (say, establish, operate and then shut down a mine – completely without any human intervention), they can also present a major threat to human society. This will need to be dealt with as the technology develops.

There is a sound philosophical foundation (and some guidance for the phase evolution and development of these systems) in European/Western philosophy, namely Koestler's "holons", Uexküll's "umwelt" and Maturana's "autopoesis". While this foundation may help to develop a theory for Autonomy of technical systems, the practical challenges will be:

(i) the development of design patterns for such systems;
(ii) the development of some kind of "operating system" that unifies the basic building blocks, in terms of the skills needed for PCA system implementation;
(iii) rules for open access to the complete software suites developed with public funding, as well as free meta-level software that *enables each individual human to (re-)gain control over the autonomous systems in as far as those decisions affect that person*.

The last point (iii) is particularly important and should be taken into account from the very beginning. It is analogous to the right to keep one's own data and privacy in cyberspace – something which was largely neglected when the Internet and later the WWW were developed, and as we can see from the recent past, is a growing problem. If Europe does not take the lead here, we will give up our cultural sovereignty completely to those who do, allowing them to become the leaders in the field – to a much larger extent than we have already done with our failure to define the rules of cyberspace in accordance with Europe's cultural heritage (even though the WWW was a development of an EU institution with public funding).

Human sovereignty over these systems is a delicate matter involving many aspects – ranging from IP issues by way of data protection, to issues of liability to free trade rules. Meeting the various requirements involved is also technologically difficult because there will be many situations in which a human being will critically dependent on an autonomous system that could very well have more information available to it than the human: think of an autonomous aircraft – what would it mean in this context that the human may always be able to assume control? These are far-reaching and highly relevant questions that need to be answered if the area of technical autonomy is to achieve user acceptance.

As we have also seen from past experience, the deployment of increasingly complex systems that people find hard to understand, can only be successful if the potential users are convinced that these systems are beneficial to them – that they improve their quality of life, that they will not harm them or the environment if they malfunction, and that they are not unreliable just because they were brought to the market too early. In other words: these systems gain user trust only (i) if their behaviour and interaction are "reasonable" by human (ethnic and cultural) standards, and (ii) if they are completely and totally reliable.

We identify the following pre-conditions for achieving **user trust**:

- *Ego-Transparency*: the system can explain itself at any given moment
- *Design-Transparency*: system behaviour is always rational and deducible)
- *Security and Guaranteed Privacy*: system detects any violation and shuts down/reconfigures if corrupted (where the definition of the "system being corrupted" is very complex if it is allowed to transform itself)

The preconditions for **total reliability** are (at least) the following:

- *Formally proven behaviour* for all possible sensor (environmental state) input envelopes
- *Autarky* (energy and information supply) – the system is always on
- *Automatic redundancy management* (the system repairs itself under all circumstances)

Autonomy: what must Europe do to lead the development?

From the description above, it has become quite obvious that we are entering a new age of technology development. If carefully thought through, this may well lead to machines that are not only autonomous during their lifetime, but that can help produce machines just like themselves, eventually even becoming independent of humans. This may transform human life in an extremely positive way, but there is clearly the danger of jeopardising it to the point of extinction. While this sounds like a potential threat, one may recall that our current societies are already completely dependent on technology for our survival, whether we like it or not.

We have to concede that Europe has largely failed to take the lead in the Information Age – today's data treasures are being watched over by Google, Apple and Microsoft. But this does not mean that Europe cannot regain the lead – on the contrary.

Our vision must be to **strive for leadership in Autonomy** – if Europe wants to play a role in the future world. More to the point: **leadership in Autonomy will largely determine the importance in the future economy**!

Fortunately, Europe has a cultural, ethical and technological tradition that puts it in the pole position for the **Era of Autonomy.** European philosophers and thinkers have defined the terms and laid the groundwork, and they are still at the forefront of research in the humanities – a wealth of knowledge

in the humanities that can now be exploited for the stepwise development of a technology that promises an interaction and symbiosis between technology and humans the likes of which was seen never before.

Moreover, the EU has invested a substantial amount of funding into advanced systems and cognitive robotics research over the past ten years; an investment, which – if managed wisely – may well pay off when it comes to the introduction of PCA-systems. It is well worth mentioning here that the EU-flagship "Human Brain Project" has a strong arm in technology development (novel computing architectures and neurorobotics), and would be an ideal instrument for connecting advanced cognitive science with the development of technical artefacts.

At this point in time, three lines of action are recommended:

a) **Content/Methods**:
   - There should be a joint, large-scale development effort for **generic architectures** for autonomous system classes (derived from robotics research), ideally using the RFC-mechanism (or an emulation thereof) used for the development of the Internet by the IETF
   - There should be a working group initially summarizing the state of the art in systems design and systems engineering, and then go on to **develop radically new design methodologies and processes** with complete abstraction from hardware and interleaving design-time/run-time phases – integrating live acquired data and formal verification of non-pre-thought situation contexts, as well as "self-evolving" design of open-ended systems with principally only partial and incomplete specifications

b) **Translation/Clustering**: on the basis of the RFCs, Autonomy Design Centres, such as fortiss, can work on architectures/tools and can then be turned into full-fledged companies. The full "value chain" must be supported with networks for innovation, combining the innovation power of large companies with SMEs ,to eventually develop into the "next SAP" for Autonomy Technology

c) **Standards**: Europe must set up standards committees for the all the base technologies – including fully open but secure and trustworthy base systems. This should be done in such a way that researchers are encouraged to participate because the work will be accorded high visibility – as is the case with RFCs – and avoid the impression that this kind of standardization is a rather boring duty. On the contrary, the results of this research will be the direct basis for the work of the design centers.

Drawing on our experience with **EIT ICT Labs, ECHORD++**, and others, the key success factors for such design centers can be identified as follows:

- **Close integration with local companies** – representing both technology providers and end users
- Establishment of a **living lab** to enable companies and interested persons to experience research results and the latest technology – "live" and with the necessary advice given by experts
- **Attractive and IP-preserving infrastructure** to enable companies to integrate their technology into the living lab and to allow end users to check solutions matching their requirements, which, in turn, will attract external researchers to stay for some time at the design center
- **Funding for small projects** conducted by researchers together with industry (both end users and technology providers)
- **Support of long-term, large projects** (modeled on the EU-flagships) to build up critical mass and to give a clear indication to the community of researchers and entrepreneurs that Europe is willing to regain the leadership.

There is a lot to be done – in terms of content, structure and organization. But one thing is for certain: Autonomy is not only a highly relevant, but also a far-reaching and innovative topic that European researchers should get excited about!