



HAL
open science

HiPEAC Vision 2024 - Rationale

Marc Duranton, Paul Carpenter, Koen de Bosschere, Thomas Hoberg, Charles Robinson, Tullio Vardanega, Olivier Zendra

► **To cite this version:**

Marc Duranton, Paul Carpenter, Koen de Bosschere, Thomas Hoberg, Charles Robinson, et al..
HiPEAC Vision 2024 - Rationale. HiPEAC. , pp.226, 2024. hal-04884304

HAL Id: hal-04884304

<https://inria.hal.science/hal-04884304v1>

Submitted on 13 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

HiPEAC Vision 2024

HIGH PERFORMANCE, EDGE AND CLOUD COMPUTING



RATIONALE



Editorial board:

Marc Duranton (editor in chief), Paul Carpenter, Koen De Bosschere, Thomas Hoberg, Charles Robinson, Tullio Vardanega, Olivier Zendra.



Funded by
the European Union

This document was produced as a deliverable of the Horizon Europe HiPEAC ("High Performance, Edge And Cloud computing") CSA under grant agreement 101069836.

The editorial board is indebted to Dr Max Lemke and Jan Komarek of the Directorate-General for Communication Networks Content and Technology of the European Commission for their active support of this work.

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those either of the full HiPEAC community or of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

The large language model ChatGPT was used to summarize and enhance some of the text in this document.

Design: Eneko Illarramendi and www.magelaan.be / Cover design: Roger Castro, Monzón Studio / Cartoons: Arnout Fierens, Arnulf.be © 2024 HiPEAC

Contents

Introduction	1
--------------------	---

Leadership race on the next computing paradigm (NCP)

Introduction.....	5
<i>by Tullio Vardanega and Marc Duranton</i>	
The Spatial Web: interconnecting people, places, things and AI for a smarter world	11
<i>by Philippe Sayegh, Safae Essafi Tremblay, Dan Richardson and Chase Pletts</i>	
Bridging the stakeholder communities that produce cyber-physical systems.....	19
<i>by Charles Robinson, with representation of communities from Rajendra Akerkar, Djamila Aouada, Alessandra Bagnato, Miklós Györfi, Michael Henshaw, Peter Gorm Larsen, Carles Hernandez-Luz, Hugo Daniel Macedo, Claudio Pastrone, Peter Popon, Claudio Sassanelli, Marcus Völp, Thorsten Weyer</i>	
Open-Source AI, why it is the best way forward for Europe	37
<i>by Thomas Hoberg</i>	

Leadership race on artificial intelligence

Introduction: The very fast evolution of artificial intelligence.....	53
<i>by various AI-based tools, including ChatGPT from OpenAI</i>	
The rapid evolution of generative AI in 2023	59
<i>by various AI-based tools, including Chat-GPT from OpenAI, Marc Duranton</i>	
AI-Assisted Software Engineering (AISE).....	79
<i>by Andreas Metzger (Editor)</i>	
Challenges in EDA: from operational research techniques to Artificial Intelligence strategies for chip design	87
<i>by Alix Munier Kordon and Lilia Zaourar</i>	
State of artificial intelligence in Europe (SWOT)	95
<i>by Jack Vernon</i>	

Leadership race on new hardware

Introduction.....	113
<i>by Paul Carpenter, André Hebben and Paolo Azzoni</i>	
Here to stay: specialized and heterogeneous computing	115
<i>by Jaume Abella, Leonidas Kosmidis, Thomas Hoberg and Paul Carpenter</i>	
Qubits and pieces: Developments and Perspectives in Quantum Computing	123
<i>by Artur Garcia-Saez, Christian Gamrat, Harm Munk and Paul Carpenter</i>	
Open hardware and RISC-V: an exciting opportunity for Europe.....	131
<i>by Miquel Moretó, Osman Unsal, Adrian Cristal, Jérôme Quévremont and Gaël Blondelle</i>	

Leadership race on cybersecurity

Introduction: The race for NCP cybersecurity.....	139
<i>by Olivier Zendra and Bart Coppens</i>	
The NCP cybersecurity challenges	143
<i>by Olivier Zendra and Bart Coppens</i>	
More data for the NCP implies more privacy risks	151
<i>by Bart Coppens and Olivier Zendra</i>	
The browser: the key to your privacy on the Web.....	157
<i>by Walter Rudametkin and Olivier Zendra</i>	
DLT and IPFS Technologies are Paving the Way for the Next Paradigm in Computing	163
<i>by Emmanuel Bertin</i>	
Integrity at Every Link: A Roadmap to Trustworthy Hardware Supply Chains.....	169
<i>by Lennart M. Reimann, Dominik Sisejkovic and Rainer Leupers</i>	
Microarchitectures as Root-of- Trust in Computing Systems – Research Needs in Formal Security Analysis	179
<i>by Wolfgang Kunz and Dominik Stoffel</i>	

Leadership race on sustainability

Introduction: The race for sustainability	189
<i>by Koen De Bosschere</i>	
What does it mean to be sustainable?	191
<i>by Koen De Bosschere and Patrick Blouet</i>	
Sustainable materials and production	205
<i>by Thomas Ernst, Lars-Åke Ragnarsson and Jean-Pierre Raskin</i>	
Towards sustainable computer systems.....	213
<i>by Lieven Eeckhout</i>	

Introduction

The last few editions of the HiPEAC Vision have taken the form of a list of recommendations shaped by a series of articles, and this release is no exception. This ‘Rationale’ document is a collection of essays on different topics, authored both by experts in HiPEAC field (principally high-performance, edge and cloud computing systems), and by experts in related areas, such as the technologies that power such systems and the applications that depend on them.

A companion to the main HiPEAC Vision 2024 recommendations, this volume comprises five sections, each focusing on one of the main areas of the HiPEAC Vision 2024. These are as follows:

Leadership race on the next computing paradigm (NCP)

Merging aspects of the web, the cloud, cyber-physical systems, the internet of things, digital twins, the metaverse and artificial intelligence into a coherent continuum of computing, the ‘next computing paradigm’ (NCP) is the backbone of the HiPEAC Vision 2024.

This section sets out what the NCP is and which areas need to be developed to make it a reality, as well as exploring societal implications of some of these computing trends.

Leadership race on artificial intelligence (AI)

One of the key ingredients of the NCP, artificial intelligence (AI) has come to dominate the computing landscape in recent years. The articles in this section – some authored using the large language model ChatGPT – take stock of the state of the art and analyse Europe’s position in the market, drawing out directions in which Europe should focus in future years.

As well as considering AI as an end in itself, articles in this section also consider the future impact

of AI on software development and hardware engineering, respectively.

Leadership race on new hardware

New workloads demand new hardware, and the NCP will be a key driver of hardware innovation. With a slowing Moore’s law and the end of Dennard scaling forcing new paths to efficiency gains, with the Chips Act poised to reinvigorate the European semiconductor landscape, and with open standards lowering the barrier to entry, there are plenty of opportunities for hardware development in Europe.

Articles in the hardware section consider the key components necessary for a quantum ecosystem, show how open hardware offers an opportunity for Europe, and explain why specialization and heterogeneity in hardware are here to stay.

Leadership race on cybersecurity

Representing a shift towards even greater interconnectivity, the NCP throws up even greater cybersecurity and privacy challenges than those faced today. In addition to a greater attack surface, the cyberphysical aspect means that cyberattacks could have tangible, even life-threatening, consequences. Cybersecurity is therefore a requirement for trust, which is essential if the NCP is to be accepted.

The articles in the cybersecurity section approach the problem from different angles. In addition to setting out cybersecurity and privacy issues in the NCP, they propose potential methods and solutions for greater trust, such as a decentralized internet paradigm, formal security methods for microarchitecture, and a roadmap to secure the microelectronics supply chain, respectively.

Leadership race on sustainability

The final tranche of articles in this volume discuss the topic of sustainability, arguably the greatest challenge of the 21st century. With information and communication technology (ICT) representing an ever-increasing proportion of the world's energy consumption, while also requiring rare minerals and precious resources for their manufacture, sustainability is becoming increasingly urgent in the electronics field.

The articles in this section discuss what it means to be sustainable, consider the issue of sustainable materials and production, and set out the main concerns for sustainable computer architecture, respectively.

You can find the full suite of articles at vision.hipeac.net

Visit hipeac.net/vision for the HiPEAC Vision 2024 recommendations and further context.

The HiPEAC Vision 2024 editorial board is indebted to the authors for sharing their knowledge for the enrichment of the community.

The Next Computing Paradigm





The next computing paradigm: an introduction

by Tullio Vardanega and Marc Duranton

What will be the future of computing systems (infrastructure, software and hardware)? HiPEAC envisions *the next computing paradigm* (NCP), focusing on a seamless integration of key ingredients from various digital elements like the Web, the Cloud, Cyber-Physical Systems, the Internet of Things, digital twins, the metaverse, and Artificial Intelligence. Envisioning the NCP emphasizes the evolution towards a spatial dimension in computing, a coherent continuum of computing, intertwining the real world with the cyberworld, incorporating Generative AI, and dynamic orchestrations of resources. The aim is to create a seamless, networked cooperative structure where resources are accessed and manipulated with streamlined Web-type protocols, where programs (in fact services) and data flow smoothly onto computing resources that cooperate with each other enhancing context-awareness and efficiency in digital interactions.

Key insights

- **Integration of Digital Elements:** This dimension combines the Web, the Cloud, Cyber-Physical Systems, the Internet of Things, digital twins, the metaverse, and Artificial Intelligence, for a holistic digital experience.
- **Spatial Computing:** This dimension adds a spatial-awareness trait to digital interactions, considering physical constraints and location-dependent factors as well as time, resulting in (at least) a 4D computing paradigm.
- **Generative AI at the Edge:** This dimension embeds personalized AI engines in edge devices for dynamic, on-the-fly construction of smart ad-hoc orchestrated applications.
- **Dynamic Web Integrations:** This dimension focuses on ephemeral, context-aware aggregations based on web resources, improving accessibility and efficiency.
- **Shift Towards User-Centric Models:** This dimension moves computation closer to users or data sources to balance latency, privacy, and energy requirements.

Key recommendations

Enacting the Next Computing Paradigm vision outlined in this part of the HiPEAC vision requires a number of distinct demanding technology improvements, each of which shall pay great attention to interoperability:

- Developing stacks of 4D-aware implementation technologies capable of spatial and time-aware computing. Doing so will entail merging streamlined evolutive versions of HTTP-based REST-like web protocols with Spatial Web protocols (HSML and HSTP, OpenUSD), and augmenting them so that they can guarantee timely delivery for any granularity of time.
- Augmenting APIs with contract-based interoperable specifications that enable ephemeral (on-the-fly) compositions based on the

pairing of assume/guarantee declarations across required and provided interfaces.

- Allowing computation to move after traveling users or objects, or toward specific data sources, seeking best balance among latency, privacy, data freshness and provenance, and energy requirements. Seeking this objective will require improving WebAssembly-type technology capable of enabling sandboxed hosting and efficient (interpreted) execution of in-transit bundles of computations.
- Developing AI-powered Edge-based trustworthy (robustly loyal) orchestrators that dynamically, opportunistically, and ephemerally assemble remote APIs into ad-hoc private service compositions.

Our modern world

If we wanted to enumerate the key elements of the digital space that surrounds us as individuals, professionals, and members of the social fabric in which we live, we would likely agree on the following list of items:

- The **Web**, as the infrastructure that supports most of our activities over the Internet. The Internet, inaugurated about 40 years ago, was the primary enabler to worldwide connectivity. The Web came some 10 years afterwards, and progressively changed just about everything as far as everyone’s experience of networking goes. If we corresponded our “navigating the network” to moving around a (gigantic and virtual) building, the Internet would be its foundation, so much below ground to be invisible, and the infrastructure of the building, which holds all contents together and allow users to move conveniently around them, would be the Web. So the Web infrastructure – its protocols and way of use – is essential to where we stand today.
- The **Cloud**, probably the most impactful by-product of the Web to date, which renders each digital thing available as a web resource in an as-a-service mode. The concept of Cloud originated from the visionary realization that everything could be exposed and access as a web resource, not only static data, but also computation (apps), and computing resources (CPUs, storage, networking). Remarkably, this vision was put forward in 1969 (!) by one of the founders of the ARPANET that predated the subsequent Internet.
- The **Internet of Things**, which timidly originated from equipping non-digital items with radio-frequency identification devices that would allow them to be interrogated digitally, if only for tracking purposes. Soon after, that rudimentary concept evolved into requiring such items to become “smart”, thus, capable of sensing and actuation, and sometimes even of basic in-place processing, eventually interconnecting them with human-side devices or among themselves. The IoT has become a rich and pervasive mesh of connected digital “things”, which allows mission-specific value-added services to be provided, to various types of target groups, up to entire populations, as it would be in a “smart city”. So we can say that the digital means (protocols and software infrastructures) to get things smart and connected is a transformative convenience to lots of our daily activities, whether professional, social, or personal.
- **Cyberphysical Systems**, which can be seen as the command-and-control processing part of all sorts of articulations of IoT devices deployed into mission-critical products that help us build “intelligent” industrial and civil infrastructures. CPSs are perhaps not very conspicuous to the general population, but they are found at all places where controlled

Intrinsic to the realization of that vision was the Web infrastructure that we have come to know as “the Cloud”, thus becoming the most global and ubiquitous programming-and-execution platform that ever existed. Very much in line with the Web logic of access to the resources exposed in it, the as-a-service model that characterizes the Cloud allows tapping applications without needing to install them locally (This should be no surprise: when you “navigate” to a resource, you consume it locally to your home base, but you know very well that it stays where it was). As part of that innovation, the REST architectural style made HTTP – the principal enabler of the design of modern Web apps – the means to access and manipulate web resources programmatically in a uniform and consistent way. REST is a most natural and productive way to leverage the as-a-service style of the Cloud: with REST, higher-level services can easily be realized by orchestrating others, in a most versatile value-added way. So we can say that the HTTP-based and RESTful view of the Web infrastructure is the foundation to modern digital services and applications.

automation is needed to guarantee the delivery of critical services in transportation, health, manufacturing, and a growing number of other sectors. As CPSs control physical devices, safety concerns arise, together with security concerns, which are common to all other critical ICT infrastructures. The central tenet of modern CPSs is its holistic view of concerns, components, and implementation competences. The range of functionalities required of CPSs increasingly include Web-enabled components, which conjoins CPSs to the landscape of the next computing paradigm.



Figure 1: a cartoon-type representation of the conjoining of the elements listed above (created by Dall-E on 6 November 2023).

- **Digital Twins**, which are the digital representation of real-world entities, hosted on compute infrastructures that may or may not be digitally connected to their actual counterpart. Digital Twins have a tremendous potential, for science, learning, conceiving, building, optimizing, planning, maintenance. They are essential for the 4D computing because they allow to use time as a real variable: we can see what happened in the past by simulation, finding the cause of phenomenon's, and also to make forecast by analysing the evolution in the future, therefore finding better options. Some capture the potential of Digital Twins under the umbrella term of metaverses (this is a 3D rendering of the realm of computing of the digital twins, visualized for humans). Regardless of the denomination, it is easy to see that

Digital Twins may be realized and exposed as web resources. If they are so, then they become part of the general (or specialized) Web space, thus making metaverses less secluded and self-contained and more permeable and pertinent to the vision we are discussing here.

Recently, two further important innovations have arisen, with potential to cause disruptive evolution of the landscape formed by the elements discussed above:

- **Generative AI**, a vertex of narrow (task-specific) artificial intelligence that is able to produce digital products of any sort, including computer programs and control commands, using “generative models”. Such models recognize structural and correlative patterns in training data extracted from specific target domains, and return rich digital outputs that feature similar patterns consistent with the received prompts. Current-generation models are massive, for size of data and need of training, but there are also smaller models emerging that are usable in a particular context. These needs demand massive investments for their production, which makes them especially attractive to commercial exploitation. Most evidently, however, the disruptive power of innovation carried by such models is also of crucial public interest. This concern will likely promote two parallel routes of evolution: (1) the development and preservation of (regional, national, continental) open-source publicly regulated foundation models, i.e., the general-purpose platforms that support the creation of generative AI applications; (2) the development of task-specific learned models and associated engines that can be deployed on resource-constrained devices for personal use or in industrial or civil infrastructures.

Route (2) will push generative AI to the Edge, and have it render personalized services. In previous editions of the HiPEAC Vision we have discussed of such AI-assisted personalized services as Guardian Angel and Digel, short for Digital Genius Loci (Duranton, 2023). This section of the HiPEAC vision builds on the vision presented in those documents.

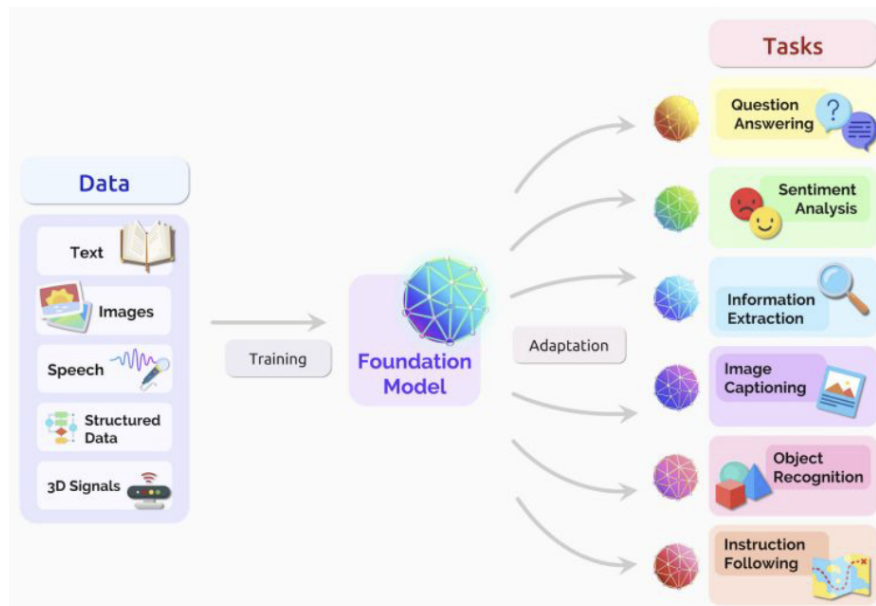


Figure 2: a view of Foundation Models as the next step from Large Language Models, and as enablers to task-specific models (source: "On the Opportunities and Risks of Foundation Models", Stanford University).

- The **Continuum of Computing**, as the digital integration of all elements listed above into a seamless networked platform where:
 - All available resources, regardless of their place of residence, are exposed as as-a-service web resources and accessed using streamlined Web-level protocols;
 - individual application services can be federated dynamically into ephemeral aggregations orchestrated into RESTful workflows originated at any point of the continuum, possibly constructed by task-specific edge-device-friendly generative AI engines;
 - the execution of the parts of those orchestrations will no longer be bound to a host device but may be able to move opportunistically from handheld or Edge devices to the centre of the Cloud, seeking to balance latency, privacy, data freshness, energy efficiency, and location requirements.

The latter prediction perhaps requires some rationale. It is vastly acknowledged that the Cloud-centric mode of service delivery causes all sorts and quantities of data to be transported from their source to the centre of the Cloud, where most application services reside, for specialized processing. This transfer is energetically costly, it is privacy-threatening, and it incurs delivery latencies that may be from fastidious to untenable for time-sensitive services. The natural countermeasure would be to have applications (or parts of them,

including their orchestration logic) travel toward the data sources of interest, instead of the opposite. As value-added applications will likely be ephemeral and opportunistic, it is impractical that they be recompiled for their destination target and equally undesirable that their ability to travel be constrained by compatibility constraints. This observation carries the implication that the execution environment at any compute node of the continuum that wishes to be part of that infrastructure be able to host and execute these application components.



Figure 3: a scene from the HiPEAC Comic Book published in 2019 (see: <https://www.hipeac.net/media/public/files/46/7/HiPEAC-2019-Comic-Book.pdf>), which evoked the Guardian Angel concept as it was described in [1].



Figure 4: How artificial intelligence (Dall-E 3) sees the continuum of computing

Key implications

The anticipated convergence of all the elements listed above will give rise to what the HiPEAC Vision 2024 terms “the **Next Computing Paradigm**”, NCP. The key tenets of the NCP vision postulate:

- The integration of the “web of humans” with the “web of machines”, where all the digital resources represented in that integration expose as-a-service interfaces that can be accessed, manipulated, and aggregated using Web-type protocols. To this end, such protocols will have to be maximally streamlined to become sustainable for use with all types of compute devices.
 - This direction of evolution will require the specification capabilities of the interface points for such Web-type protocols to be augmented to capture an increasing range of non-functional requirements (energy, latency, provenance, service level, etc.)
- The impetuous emergence of a spatial (and time) dimension to the next-generation web-inspired platform evoked above, which will be crucial to warrant context-awareness in the regard of physical constraints, location-dependent rules (e.g., norms and legislation), local knowledge. The spatial augmentation of Web-type protocols will:
 - Require a standard language to encode properties of physical objects and spaces, logical concepts and allowable activities associated with them.
 - Require a suite of standard protocols to expose contract-based interfaces associated to zones and objects, and to support credentialed requests and interrogations on them.

- Equally apply to mobile computation (execution bundles that may move for and during execution) and to mobile devices (where executions are statically bound to a moving host), unmanned aerial vehicles, and other semi-autonomous transportation vehicles.

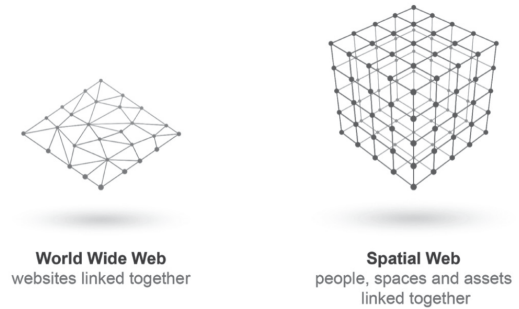


Figure 4: a view of the Spatial Web contrasted with the Web as it currently is. Excerpted from “The Spatial Web”, by G. René and D. Mapes, 2019, page 35.

- The envisioned spatial computing will be CPS-like (operating with and for physical systems, coping with time constraints of the real world), swarm-like (supporting opportunistic dynamic and mobile aggregations of compute nodes within variable-size logical regions), 4D-enabled (fit for extended reality, spatial digital twins involved in time-sensitive operation).
- The embedding of Generative AI engines (models and prompt handlers) in Edge devices, whether handheld or deployed in industrial or civil installations to provide for the on-the-fly construction of ad-hoc applications expressed as dynamic, opportunities, ephemeral, smart orchestrations of calls to as-a-service interfaces exposed in the logical or physical regions of interest. Interestingly, such regions can well be temporary self-sufficient federated clusters of Edge devices that may even occasionally happen to be partitioned from the Internet. The prompts that will trigger creation, deployment, and execution of these dynamic orchestrations will use any conceivable “natural” interfaces, including voice for humans, and video imaging for automated requestors.

Conclusion

The document presents a transformative vision for the future of computing, emphasizing the integration of digital and physical realms through advanced web protocols, spatial awareness, and Generative AI. This paradigm shift aims to create a more efficient, context-aware, and user-centric digital ecosystem, paving the way for innovative applications and services in various sectors. The success of this vision hinges on technological advancements and collaborative efforts across multiple domains.



Figure 5: *The Next Computing Paradigm*, hallucinated by Dall-E 3.

References

Duranton, T. V. (2023). “Digels”, digital genius loci engines to guide and protect users in the “next web”. In M. D. al. (Ed.), *HiPEAC Vision 2023*, (pp. 18-21).

Tullio Vardanega is an associate professor in the Department of Mathematics at the University of Padova, Italy.

Marc Duranton is a researcher in the research and technology department at CEA (the French Atomic Energy Commission) and the coordinator of the HiPEAC Vision 2023.



We stand on the brink of a technological revolution where the web as we know it will break out from behind the screen and exist in the physical world along with us. Sometimes referred to as Web 3.0, the spatial web will seamlessly blend our physical and digital worlds together.

The Spatial Web: interconnecting people, places, things and AI for a smarter world

by Philippe Sayegh, Safae Essafi Tremblay, Dan Richardson and Chase Pletts

The Spatial Web is a concept in the evolution of the internet that envisions a multi-dimensional online space intertwined with the physical world. Unlike the traditional two-dimensional web, the Spatial Web integrates digital information with physical locations and objects, creating a seamless blend between the digital realm of ones and zeroes, and the physical realm of places and spaces. This unifying web marks the evolution from a network of pages to one of spaces – cyber-physical and conceptual alike – and will interconnect activities, people, places and things, as well as AI to form a smarter world.

AI will join with existing and new Cyber Physical Systems (CPS), e.g., sensors and actuators, IoT devices and appliances, and autonomous vehicles, to become Autonomous Intelligent Systems (AIS), stemming from the fusion of AI, CPS and the IoT.

This paradigm shift will impact massively on the fabric of our professional, social, and personal life, creating smarter urban environments, advancing personalized healthcare and immersive educational experiences. Embedded computing devices will enable feedback loops where physical processes affect computational learning and vice versa.

The Spatial Web brings with it the possibility of creating a smarter world with new realms of possibility for individuals, organizations, and civilization as a whole.

Key insights

- We're entering a new era of the internet: the “Spatial Web” – where the digital and physical worlds merge seamlessly, where compute and data knowledge will come out of our screens and into the world. This Spatial Web is sometimes called web 3.0, industry 4.0, the metaverse, or society 5.0, depending on your vantage point. It integrates AI with Cyber Physical Systems, creating automated and autonomous Intelligent Systems. It will revolutionize everything, from urban living to healthcare and education, making the world smarter and rich with new possibilities.
- This next web will be a network of distributed intelligent agents (both human and machines)

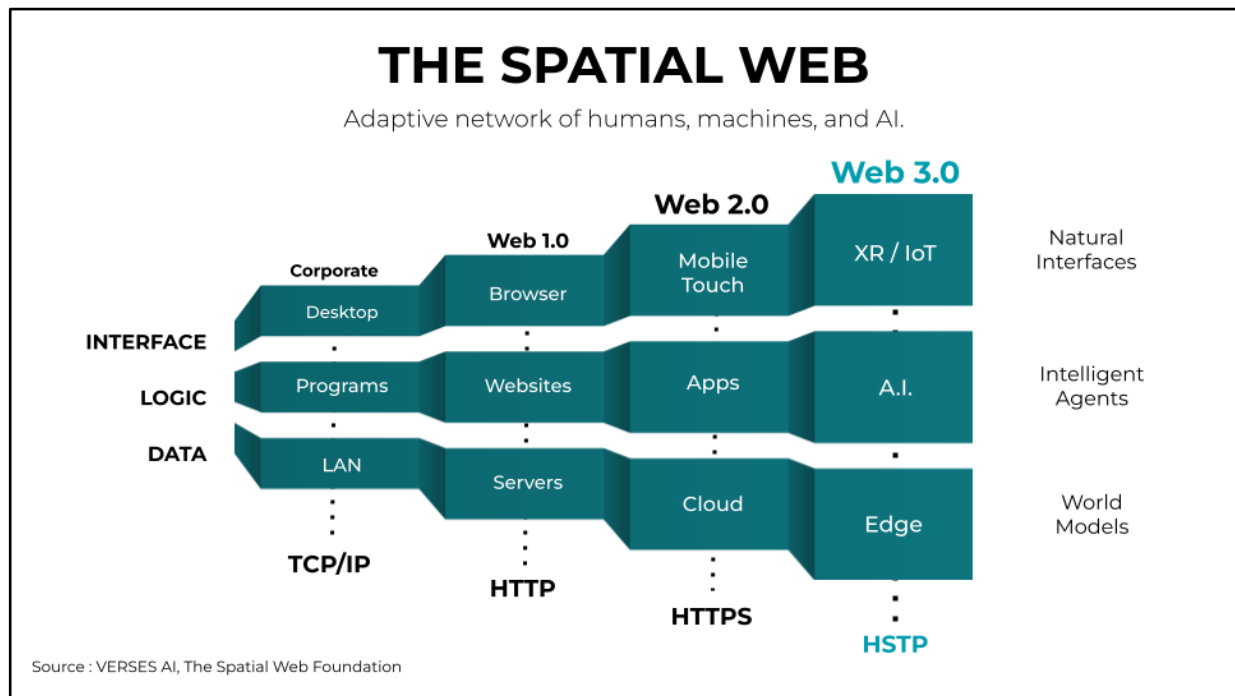


Figure 1. The Spatial Web concept in relation to its predecessors.

working together. To ensure that these autonomous systems understand, operate and connect with each other within the boundaries of safety, privacy, law, and ethics, new types of orchestration and governance will be needed.

- The underlying infrastructure will need to be augmented and will require new socio-technical standards that are designed to provide governance capabilities, transparency, and auditability for autonomous agents and ecosystems of agents.
- Furthermore, a shift in AI methodologies towards an approach based on Active Inference will enable AI that is transparent and explainable. AI agents that leverage the Active Inference framework are able to continuously sense, understand, predict, and act. This ongoing cycle produces AI that can adapt and evolve over time. They are explainable and self-learning by design.

Key recommendations

To propel the Spatial Web towards its maximal potential, we propose the following strategic initiatives:

- **World Model Creation:** Develop models based on standards and spatial and multidimensional data to simulate impactful scenarios.
- **Promotion of socio-technical standards:** Advocate for standardization in adaptive computing and metadata.
- **Upskilling:** Encourage talents to master the implementation and methodology of spatial web standards.
- **Join Collaborative Bodies:** Participate in groups like the Spatial Web Foundation and the IEEE Spatial Web Working Group (P2874).
- **Awareness and Collaboration:** Engage in hackathons and competence networks to drive adoption of Active Inference based AI.
- **Funding:** Allocate funds to support startups and applications using these standards and active inference based AI.
- **Industry-wide Implementation:** Collaboratively work on broad-scale implementation of socio-technical standards.

Challenges and requirements

With all this power for change comes great responsibility. The convergence of exponential technologies has the potential to test civilization as much as it can help it. This shift necessitates responsible stewardship and serious ethical consideration. As industries are transformed by AI and our quality of life improves, it will be critical that governance keeps pace with innovation.

The emergence of Autonomous Intelligent Systems (AIS) introduces an entirely new set of challenges. Advanced entities, such as autonomous drones, sophisticated manufacturing robots, and interactive companion devices, will operate autonomously, learn from their surroundings, and effect tangible changes in the real world. Ensuring universal interoperability becomes essential as the current 8 billion humans and 30 billion devices are expected to come online by 2030 to the Spatial Web.

Promoting this vision begs the following questions:

- How do we get all these diverse systems to talk to each other?
- And once they can talk to each other, how do we govern these systems that are on the path to become self-governing?

The answer lies in the foundation of the Spatial Web, i.e., an augmentation of the current internet infrastructure with new standards, a new approach to modeling data and to AI. The new standard infrastructure will become the fabric that we use to connect to AI, and that AI agents will use to connect with each other. These innovations will allow controlling how knowledge is structured and how information is shared on the network, also making it possible to build governance directly into the web itself. For this to happen, we need to upgrade the standards and protocols that are the backbone of the current Web 2.0.

Socio-technical Standards enable shared understanding between humans and machines

The web as we know it today runs on a suite of technical standards, where HTML and HTTP have become predominant. The web technologies were not explicitly tailored to handle the demands for transactions, interoperability, security, and privacy, of contemporary complex systems and the connected smart world. However, they were open by design, which allowed for sufficient though laborious and creative adaptations.

As AI turns into an online commodity and becomes networked, the privacy and security challenges that are inherent in Web 2.0 technologies will grow exponentially. Ultimately, it may become impossible to course-correct as AI gets more and more powerful. It is therefore the case that we should turn our attention to the fundamental infrastructure of the web: the standards that define it.

Historically, society has deployed technical standards to foster safety and interoperability in the use of technology.

“Cars run on gas from any gas station, the plugs for electrical devices fit into any socket, and smartphones connect to anything equipped with Bluetooth. All of these conveniences depend on technical standards, the silent and often forgotten foundations of technological societies.”

– *The New York Times*¹

Considering the power of AI to alter virtually every sector of the world economy, technical standards aren't enough. A new generation of web standards will also need to address social requirements around transparency, explainability, accountability, safety, and other societal or human-centered values.

A hybrid approach of socio-technical standards can bridge the gap between technology and society. Socio-technical standards could enable AI and AIS to be technically sound, socially beneficial, safe, compliant with laws, and able to be aligned with societal norms and values.

¹ <http://tinyurl.com/mrxnab5c>

In 2020, The IEEE P2874 Spatial Web Working Group was formed to lead the development of socio-technical standards for AI and AIS alignment, interoperability, and governance. These standards are informed by IEEE's Ethically-Aligned Design P7000 Series of standards that address human rights, well-being, accountability, and transparency for AI and AIS.

The IEEE P2874's Spatial Web Standards are being developed to address the following:

1. Shared understanding of meaning and context between humans and AIs.
2. Explainability of AI systems, enabled by the explicit modeling of their decision-making processes.
3. Interoperability of data and models that enable universal interaction and collaboration across organizations, networks, and borders.
4. Compliance features that are built to adapt by design with diverse local, regional, national, and international regulatory demands, cultural norms, and ethics.
5. Authentication and credentialing, driving compliance and control over critical activities, with privacy, security, and explainability built-in by design as well.

These standards lay the foundations for the efficient integration and adoption of AI technologies while minimizing the risk inherent in AI. In the sequel, we highlight a few essential components of the Spatial Web Standards.

Socio-technical Standards enable comprehensive world modeling

World modeling in AI involves creating internal representations of the external environment, utilizing abstract symbols to understand and interact effectively. However, this process encounters a crucial challenge known as the grounding problem. This challenge emerges when translating symbolic representations into a meaningful reflection of real-world entities, requiring a bridge between the abstract and the concrete. Successful world modeling addresses the grounding problem by integrating sensory input and learning from real-world interactions. The resolution of the grounding problem enhances the accuracy and context-awareness of AI systems, enabling them

to navigate diverse environments with a deeper understanding.

The successful implementation of AIS is therefore dependent on their ability to create comprehensive and dynamic world models. AI and AIS systems will need hyperdimensional world modeling to enhance performance and explainability. These systems must be adept at understanding and interpreting complex models of the world from as many perspectives and sensory inputs as required by the problem they are trying to solve or the activity they are trying to predict and optimize. For IoT and cyber-physical systems to stay pertinent and adjust to shifting use cases and scenarios, data must be integrated within a broad world model.

World modeling is multi-dimensional. It encapsulates identities, activities, environments, policies and credentials, which need to be expressed in a coherent and shared manner in different contexts:

- Semantic (meaning and logic)
- Spatial (physical and situational)
- Societal (values and value)
- Systems (networks and ecosystems)

Comprehensive world modeling needs to:

- Be stateful
- Be multi-modal / multi-dimensional
- Be interpretable and actionable by machines
- Be shareable between heterogeneous networks, devices and applications, and humans
- Maintain coherence over time and space for all the actors/edges involved in a use case

The Spatial Web socio-technical standards, Hyper-Spatial Modeling Language (HSML) and Hyper-Spatial Transaction Protocol (HSTP), enable world modeling by structuring spatial information and securing efficient transactions.

HSML

Hyper-Spatial Modeling Language is a knowledge modeling language that enables systems to encode properties of physical objects, logical concepts,

and contextual activities. HSML facilitates multi-modal world modeling and knowledge sharing among machines and humans, encompassing ethical, moral, economic, and societal considerations. HSML models relationships and activities, addressing the Who, What, When, Where, How, and Why.

HSML allows for the detailed description of entities and their relationships with other entities within the Spatial Web. It serves as a modeling language and semantic data ontology schema, crucial for creating complex and accurate models of spatial environments and contracts. By providing a framework for encoding these models in a way that is both human- and machine-readable, HSML facilitates the construction of dynamic, interactive world models.

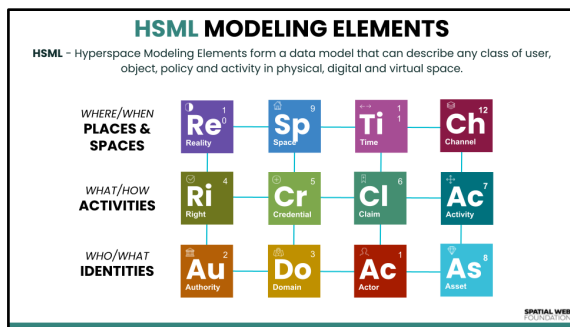


Figure 2. Modelling elements of HSML.

HSTP

Hyper-Spatial Transaction Protocol provides the methods for passing HSML messages in the Spatial Web. It provides a universal, secure, and verifiable protocol for communication between digital or physical systems, ensuring seamless interaction and cooperation between diverse AI systems. It incorporates a zero-trust architecture and strict authentication measures for secure data exchange and control over AI operations.

HSTP manages the transactional aspect of models of spatial environments and contracts. It is designed to support automated contracting, ensuring decentralized, secure, and privacy-respecting interactions within the Spatial Web. By providing APIs for distributed computing platforms, HSTP enables smooth and secure exchange of information and resources within the modeled world, thereby supporting dynamic interactions and operations in world modeling.

HSTP's zero-trust architecture ensures that data sharing across environments is done so with security and privacy principles embedded by design, in particular as it mandates verifiable credentials for any interaction between systems. This rigorous credential-based authentication process is particularly crucial for AI activities, as it allows them to operate within a secure and compliant framework,

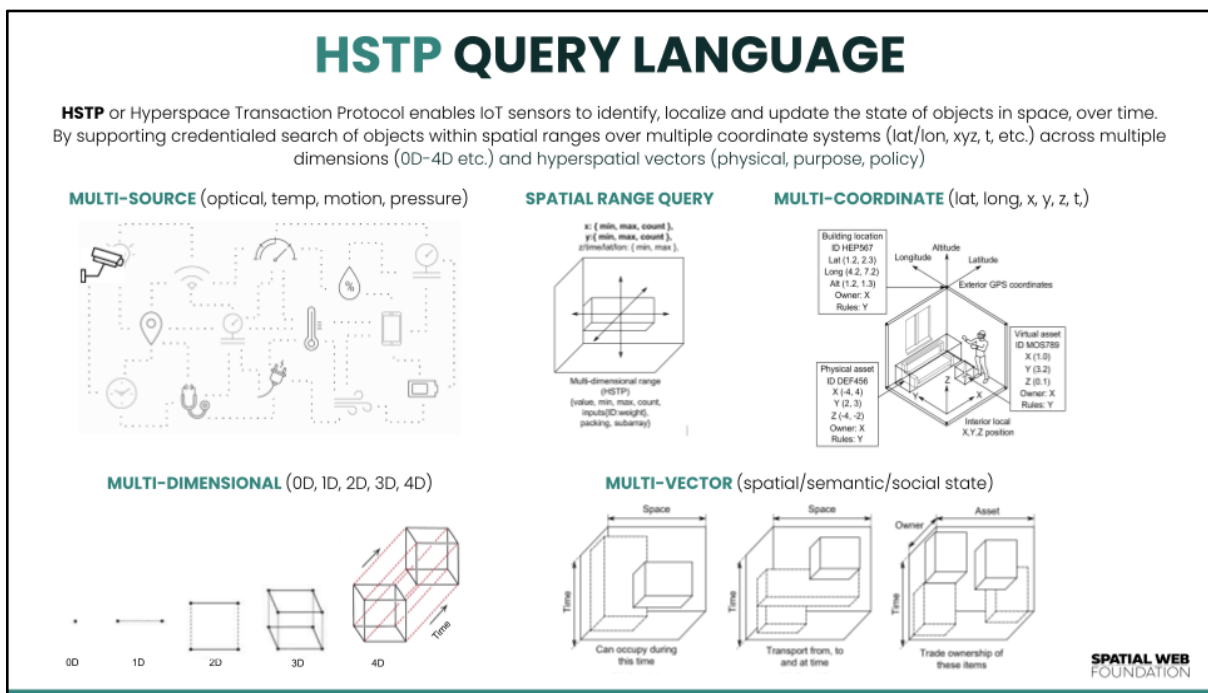


Figure 3. A simplified view of the HSTP query language.

protecting against unauthorized access, ensuring the integrity of data and operations, and significantly enhancing the security and trustworthiness of all operations across the Spatial Web.

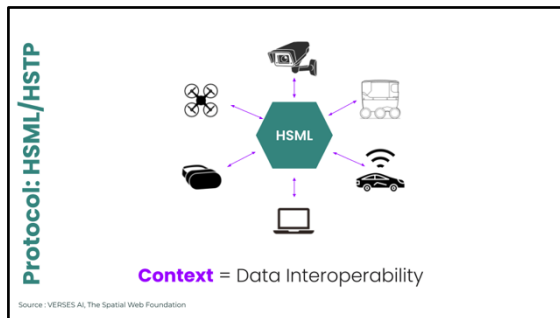


Figure 4. HSML over HSTP.

In contrast to the open structure of the Internet and the World Wide Web, the Spatial Web, based on HSTP, is designed as a permissioned network by design. This fundamental shift in architecture not only enhances security but also increases the reliability and predictability of AI operations within this environment.

A new Approach to AI

The field of AI is at a critical juncture where traditional methods, often based on narrow, task-specific algorithms, are reaching their limits in terms of adaptability, generalizability, and understanding complex, real-world environments. This limitation calls for a new approach that can bridge the gap between highly specialized AI and the more versatile, adaptive intelligence seen in natural systems. Active Inference emerges as a promising answer to this challenge.

Active Inference

The framework of Active Inference decodes biological intelligence by analyzing how the human brain creates mental models and makes predictions based on those models. VERSES is now applying this framework to the fields of computer science and AI.

Originally developed by Karl Friston—most highly cited neuroscientist², theoretician at University College London, and Chief Science Officer at VERSES—active inference has the potential to completely transform the field of artificial intelligence by creating intelligent agents that can model the world, and use those models to think, plan, predict, and act.

Active inference defines in mathematical terms the process by which agents, whether living organisms or digital systems, learn by interacting with their environment. It posits that all intelligent agents are fundamentally engaged in minimizing the uncertainty between expected sensory inputs and actual sensory inputs. Agents make predictions about the world and then act to make those predictions come true. The goal is to reduce the level of surprise an agent experiences. The less uncertainty an agent must navigate, the better its chance of survival.

Active inference involves creating an internal representation of the external world. This world model enables active inference-based agents to make predictions about the causes of sensory inputs and the likely outcomes of actions.

As the agent takes action in the world, it may learn something new about its environment. It can then update its world model with this new information. Agents can also access a shared world model that is continuously updated by multiple agents, resulting in a world model that is shared and always up to date.

For AI and robotics, active inference offers a blueprint for creating systems that can autonomously learn and make informed decisions. These agents use predictions based on an always-up-to-date world model to guide their behavior, constantly adjusting their actions based on new data from the environment. This creates a feedback loop where the AI's actions are both informed by and inform its predictions, leading to a self-correcting system that can become more sophisticated over time.

The shift towards self-learning and agentic AI necessarily poses crucial regulatory questions – How do we effectively regulate a system inherently designed for self-regulation? The evolving nature of AIS and its potential autonomy raise urgent

considerations for regulatory frameworks, prompting a need to strike a balance between fostering innovation and ensuring responsible governance.

Governance and Regulations

Historically, laws have primarily focused on human-to-human interactions, where actors and subjects are humans or human-controlled entities. With the arrival of AI, however, there has been a shift in focus to human-to-AI interactions. This includes issues such as data privacy, algorithmic bias, intellectual property rights in AI-generated content, and questions around liability for decisions made by AI systems. Laws are being updated and new ones created to address these unique challenges, where the lines between the creator (human) and the creation (AI) are often blurred.

Looking ahead, there's an anticipation that the legal system will need to evolve further to govern AI-to-AI interactions. This emerging field is likely to present unprecedented challenges. Key issues may include the autonomous decision-making by AI entities, interactions between different AI systems without direct human oversight, and the consequences of these interactions. For instance, two AI systems might negotiate contracts, conduct transactions, or even engage in conflict resolution without human intervention.

For AI systems to understand and apply laws, these laws must be converted into a format that machines can understand and process. This involves translating legal texts into structured data that can be easily understood in all contextual dimensions by computer algorithms. This would mean coding laws in a way that captures their essence and directives without ambiguity, which is a significant challenge given the complexity and nuanced nature of legal language. Beyond just being readable, laws need to be interpretable by AI. This means that the AI must be able to understand the intent, context, and application of the law. Developing a universally accepted socio-technical standard for how laws are encoded is therefore crucial. This ensures consistency in how different AI systems interpret and apply the law. Without standardization, there could be significant discrepancies in legal interpretations, leading to unpredictability and potential injustices. Moreover, laws evolve over time, responding to societal changes, new understandings, and precedents. AI systems will need to be adaptable to these changes, requiring

mechanisms for continuous learning and updating of legal knowledge bases.

Additionally, machine-readable and interpretable laws must be accessible to those who are subject to them, including humans. This means that while laws need to be encoded for machines, they also need to be understandable by humans in a transparent manner, ensuring that the legal process remains open and fair. These laws also need to be explainable, in order to respond to auditability and liability concerns. It is crucial that the rationale behind automated legal decisions is clearly outlined and can be scrutinized. This ensures that in cases where disputes arise, or errors occur, there is a traceable decision-making process. This not only aids in holding systems and their creators accountable but also fosters trust in the technology by demonstrating that decisions are made based on logical and fair principles, and that there are mechanisms in place to rectify any mistakes or biases.

AIS International Rating System (AIRS)

Intelligent machines may need to operate optimally across a range of governance frameworks, from centralized to federated to distributed. Standards that facilitate interoperability across the spectrum will be essential.

The chart below illustrates how AI systems that become increasingly intelligent gain the potential for greater autonomy, which is reflected in the corresponding governance framework that becomes available, along with all other governance frameworks that came before it.

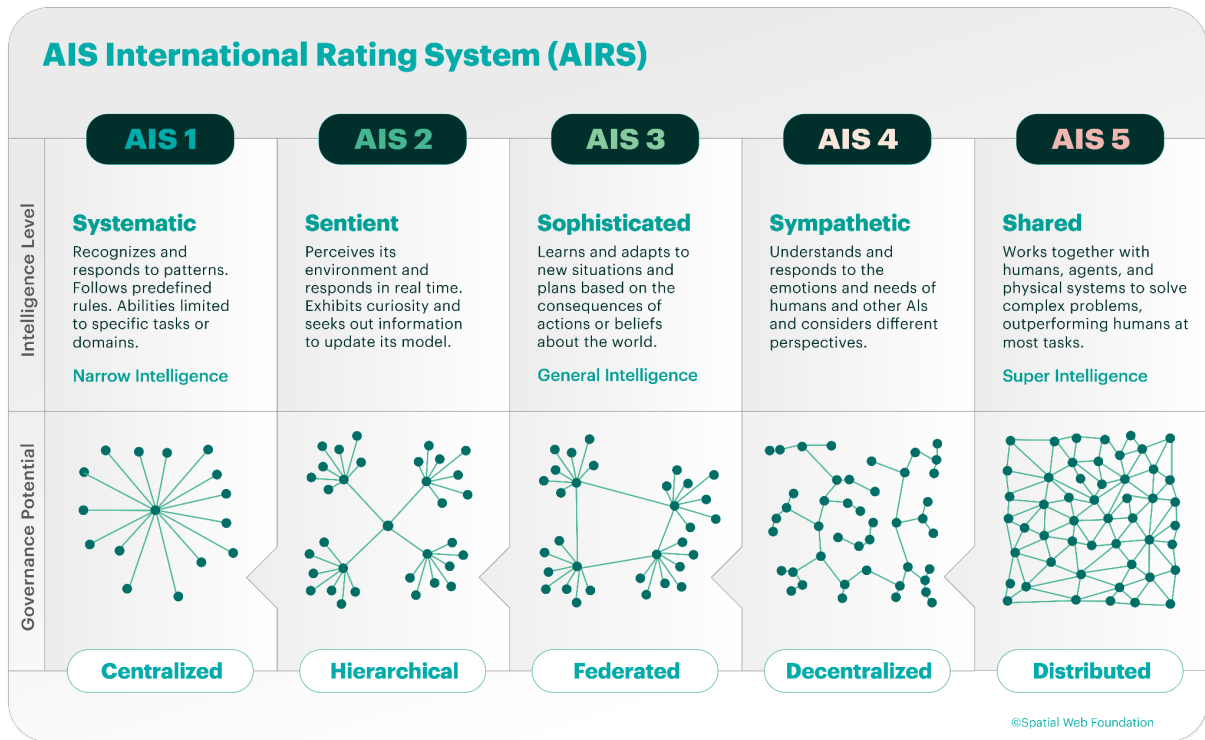


Figure 4: AIRS Chart sourced from: *The Future of Global AI Governance*

Philippe Sayegh is chief adoption officer at VERSES.ai.

Safae Essafi Tremblay is senior grant manager and researcher at VERSES.ai.

Dan Richardson is director of market analysis at the Spatial Web Foundation.

Chase Pletts is senior editor at VERSES.ai.



The Next Computing Paradigm will draw upon the continued integration advances for the development of future large-scale safety-critical systems, involving many technology and influencer communities. Novel approaches and tools will be required to tackle the increasingly multi-dimensional challenges between the communities to benefit these future systems, especially their adaptability to new technologies.

Bridging the stakeholder communities that produce cyber-physical systems

by Charles Robinson, with representation of communities from Rajendra Akerkar, Djamila Aouada, Alessandra Bagnato, Miklós Györffi, Michael Henshaw, Peter Gorm Larsen, Carles Hernandez Luz, Hugo Daniel Macedo, Claudio Pastrone, Peter Popov, Claudio Sassanelli, Marcus Völp, Thorsten Weyer ¹

Today, there are many communities involved in the creation of large-scale safety-critical systems, which are used in domains including transport, health, robotics, manufacturing and, in the longer term, will be in the home, where miniaturization will play a role. These applications have the encompassing term of cyber-physical systems (CPS), of which CPS technologies consider the integrations of the system parts across the digital and real world boundaries. Thanks to great advances in connectivity (including IoT, Big Data), pattern recognition and decision making (AI) and integration technologies (CPS), with computing as a backbone, there is a convergence of technology domains and the real world towards the Next Computing Paradigm (NCP). The integration concerns typical in the realm of CPS, including certified dependability, energy management and real-time capacities, will become pervasive for NCP and tomorrow's systems.

In this article we explain that engineering for future CPS needs a centre of gravity in order to draw

these communities together. This will provide common goals around which technical advances can be aligned. Overviews of the communities involved are provided, with examples of their relevance in the creation of CPS and to some common challenges.

Advancements of aggregating technologies are multi-dimensional challenges, representing many influencing dependencies from all communities, especially at higher levels where the whole system product is drawn together. This means that, to make good progress, Europe will require new forms of coordination in order to orchestrate research and to capitalize on lessons learned related to the cumulative advances between the communities.

Key insights

- Large-scale safety-critical systems, encompass a common perspective of what is classed as

¹ For the list of contributors, see the acknowledgements

cyber-physical systems [39]. These CPS are physically interactive (high certification obligation) and increasingly collaborative (task sharing). They involve many contributor and influencer communities in their creation, who each tend to make advances in isolation. *Creating the technical bridges between these communities to channel technology development is essential for these future systems.*

- The scope for bridging the communities is wide and they need technical interfaces around which to align. Discussions have indicated self-alignment of these groups is needed through *centres of gravity, particularly on the topic of real-time safe and secure automation.*
- *A new form of research coordination is necessary to direct cumulative integrated developments from the stakeholder communities.* CPS projects should provide technical advances on researcher supports in addition to system technologies.
- The development of CPS requires a holistic approach, guided by target products, that brings together a wide range of disciplines. These should include not only functional, system and enabling technologies, but also the fields of psychology, sociology and ethnography, among others.
- Aggregating, or system technologies, have different and much longer industrial uptake lifecycles than part specific or component technologies. Research programmes mostly treat them the same and both technology types suffer. *A dedicated team from the research programmes would be very beneficial for investigating and implementing new technical capacities for multi-stakeholder complex group research*
- CPS represent a significant part of national infrastructures and where lie some of the most devious and complex research integration challenges to be solved. Infrastructure represent an important means of market capture and thus sovereignty, whilst system architecture is a determining factor of technology uptake and the green transition. National funding for infrastructure stability and adaptability, particularly investing in system thinking/interdisciplinarity will play a critical role for Europe on the world stage. This is not only in terms of economic stability and productivity but also moving to a culture that ensures also a functioning natural world for future generations.

Key recommendations

- Develop new support capacities to carry out cross-domain research on multi-dimensional critical complex group challenges. Draw on best practices for systems thinking\interdisciplinarity to develop methodologies and supportive tooling that will act as the technical bridges between communities contributing to large-scale safety-critical systems. It is relevant to considering other interdisciplinary fields treating complex systems such as governance of climate change and world food security.
- Motivate convergence through centers of gravity, including advancing an integrated view in real-time, safe and secure automation, a common interest for communities contributing to large-scale safety-critical systems.
- Build up research orchestration techniques for coordinating cross-community research to future large-scale safety-critical systems. This also calls for projects tackling challenges that are common across the contributing communities.
- Establish a research instrument, transversal to European research programmes with stable financing, for advancing aggregating technologies in particular and the technology uptake by large-scale safety-critical systems in general.

Introduction and new cross-community development approaches

In order to manage large complex problems, people break them down into parts. It is for this reason that, from the technology point of view, there are many contributing and influencing communities involved in the creation of future safety-critical products. Of course, the parts subsequently need to be assembled together in order to address the initial complex problem. For the same reasons, the various technological contributions for future large-scale safety-critical systems require layered aggregation in order to achieve these physically interactive and collaborating systems.

This means that there are significant, multi-dimensional influences across CPS communities, which contribute to our ability to transfer technology to industry. It also poses challenges for assuring CPS,

which should be based on sound methods of justifying that a CPS is fit for purpose and that all risks of using it are adequately addressed, notwithstanding the complexity and the heterogeneity of the CPS components and of the communities of stakeholders involved. For instance, assurance alone historically has relied heavily on *expert judgement* and this worked. Recently the complexity of the CPS that we try to assure as safe has increased very dramatically. In these circumstances, expert judgement based on previous experience alone becomes problematic. The current trend (see Assurance 2.0 [40]) is that we need new methods for building assurance cases which should rely on formal methods and also on *automation* to process complex arguments on which assurance is built.

For the purposes of this article, we take CPS in the context of an application; that is to say, the term could be replaced directly with an example CPS application such as railway transport, an autonomous vehicle, or satellite constellations. In this framing, CPS therefore represent *physically interactive and collaborating systems* that are present in many domains including transport, health and manufacturing. (For an in-depth definition of CPS, see, for instance [39]).

Communities involved in CPS, discussed in the subsequent section, range from providers of a) functional properties including sensing, physical action, communication, energy provision, processing and coordinated collaboration to b) system-level engineering including properties like safety and performance specifications, managing customer requirements, architecting, system verification and validation, mechanical engineering and control engineering. There are technology support communities providing c) enabling technology domains like the Internet of Things (IoT), Systems of Systems, Big Data, Artificial Intelligence/machine learning and High-Performance Computing. Finally, there are the influencing communities from d) the production environment, with enterprise processes and product line, and e) the market, such as regulation and current and future needs of society.

These communities have tended to transfer technology as a one-to-one mapping with products. However, to respond the challenges of future CPS and to enhance technology transfer, they will need to take relations with the other contributing communities increasingly into account. While the challenges and importance of advancing aggregation techniques are discussed later, there also needs to

be a common focal point from which one community can interact with any of the other communities. This point should provide a common interest based on the physical challenges of these systems. *Discussions have proposed this centre of gravity to be real-time, safe and secure automation* of CPS development and operation.

Research on CPS should seek to enhance the interrelations and automation of these three dependability properties, i.e., real-time, safety and security. They are goals that must be achieved at a global level when all the technologies are combined. As an example, each piece of hardware has an impact on the energy consumption of the whole system. Similarly, individual software and hardware components can jeopardize safety if they fail naturally or due to a security breach. These goals can also be variable and related to environmental conditions, such as a train reducing its speed (performance) in response to heavy showers (to maintain safety).

Hence for technologies to be accepted in these systems, they must guarantee these dependability properties, i.e., they must comply with the safety and security constraints of a product in the intended operational environment and not violate the constraints on real-time responses. This means that the easier it is to couple your technology with these system constraints (through automation), the easier it becomes to adjust it to the system (or adjust the system for new technologies).

It is usually the case that, in order to add new technologies to a CPS, the whole system requires recertification. This can be prohibitively expensive without sufficient automated information about the impact of the new technologies on these dependability properties – and particularly the interrelations of those properties. Take systems certified, for instance, against an extreme earthquake occurring every 1000 years (for safety), such as a nuclear plant: in this case, the safety experts currently would prefer no new technologies or patches for security to be added to these systems due to the certification costs.

As a result, historically, interrelations between system properties have been limited to minimize complexity, but the current need for adaptability (to new technologies, to environmental or internal changes) requires this design mindset to be readdressed. So in summary, a centre of gravity, as shown in Figure 1, will provide a useful point to

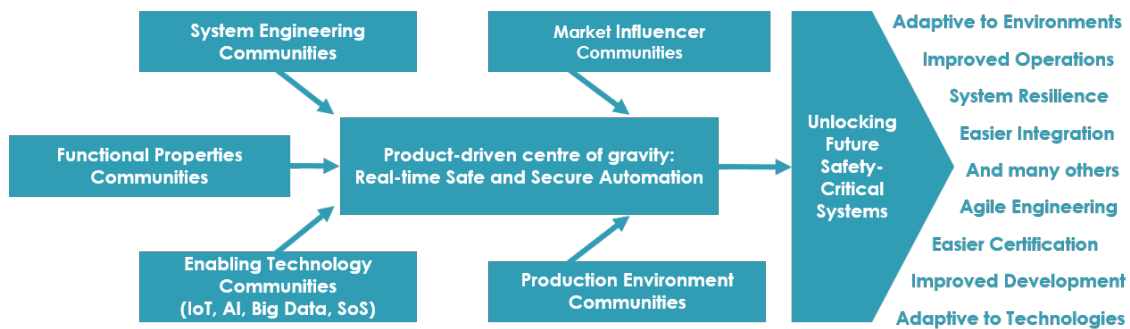


Figure 1: The stakeholder communities for creating CPS.

channel us towards more impactful research advances for these future large-scale safety-critical systems.

While the management of trade-offs between the system properties of performance, safety and security is an established skill in system development, it still remains very much a manual and qualitative process and one that is based on prior experience. It remains to this day very much a bottleneck and is holding back the communities contributing to CPS development from ensuring that advances in areas such as trust in artificial intelligence (AI) are applied to CPS.

System-level engineering for CPS is therefore in need of transformative automation. Fortunately, automation between system-level dependability properties can rely on a number of decades of research in techniques [1], some of which have already been applied in industry but are generally in need of new approaches for technology transfer. Such approaches are included in the coordination suggestions for research orchestration described later in this article.

Of course, current pressures for industry to find advanced solutions for managing system property trade-offs are also driving the search for automated coupling. As examples of some initiatives, the UK Research Institute in Trustworthy Interconnected Cyber Physical Systems (RITICS) [2] involves dozens of UK universities and industrial collaborators. Topics include safety, security and autonomous systems. Relating to autonomous vehicles, the Intel Research Collaborative Institute of Safety of Autonomous Cars (ICRI-SAVe [3]), deserves a mention as a vibrant community. Likewise, the VVM project (Verification and Validation Methods, <https://www.vvm-projekt.de/>) built up a sizeable research community addressing

the challenges of “safety verification of automated vehicles ... on driving functions up to full automation of vehicles (SAE Level 4 and 5)”.

Many industries are actively looking for solutions to manage the performance, safety and security of their products, including large enterprise like Siemens, Thales and AVL, who have been forming combined safety-security teams. The challenge also affects small and medium-sized enterprises (SMEs) in their products and services. This recent momentum has visibility, for example, in the Ada and IEEE conferences, in the IET code of practice on cybersecurity and safety [4], in recent large research collaborations including MERgE [5], SeSaMo [6] and AQUAS [7], and in co-engineering discussions.

Overview of stakeholder communities for creating and advancing CPS

We now provide overviews of the five communities, indicated in the previous figure, which are involved in creating CPS. We give descriptions and examples of their relevance to CPS as well as their relation to cross-community challenges for future development. These include embedded computing as a CPS backbone, system decentralization and decomposability, and physical collaborations with people.

Functional-property communities

CPS functional properties must address aspects that cover sensing, actuation, communication, energy provision, processing and coordinated collaboration. Such properties are key characteristics of these systems, with actors in specific communities researching and developing the different components.

The relevance of functional properties becomes more evident when considering novel and innovative advanced applications that are being progressively adopted in several large-scale, safety-critical domains, such as industrial automation, transportation, smart cities, critical infrastructures, space, etc. Some examples can be found in H2020 projects such as CPSwarm [8], or Chips Joint Undertaking such as AIDOaRT project [36] and other CPS cluster initiatives.

Industry-driven needs and the well-established nature of general research communities in the CPS domain mean that it is feasible to envision projects that might prototype concepts such as swarms of unmanned aerial vehicles and rovers supporting safety and security operations; swarm of unmanned aerial vehicles and possibly ground robots supporting critical infrastructure management; swarms of automated ground robots that collaboratively support humans in logistic operations within a smart warehouse or in last mile delivery operations within a smart city; or enhanced and dynamic platooning applications for autonomous freight vehicles. Currently, the development of such applications cannot leverage a simple plug-and-play integration of the various technologies entailed, given the complexity of managing teams of systems and humans in evolving and dynamic scenarios with emergent properties.

Therefore, to properly combine and integrate the different technology building blocks required, the various ‘functional properties communities’ have to be properly engaged. Experts from the functional property communities will need to work with other actors with collaborative systems competence. Moreover, while the increased adoption of CPS has resulted in the maturation of solutions for CPS development, a single consistent science for future CPS has not yet been consolidated. Few functional properties community members have already started working alongside other communities on a connective framework e.g., using modeling, design/development tools and methodologies, deployment solutions, monitoring and controlling solutions for large-scale challenges. In this context, model-centric approaches have clear relevance for facilitating collaboration between experts from different sectors and thus enabling the definition, composition, verification and simulation of collaborative, autonomous CPS.

For these reasons, it is important for future CPS to be considered not only from the technology perspective but also as an application domain

where the technology of the functional property communities plays a role for aggregation of CPS-related research. To promote this, closer and wider collaboration is needed within the communities, along with new research initiatives. Understanding the nature of this aggregation from the bottom up and top down is important for driving the communities towards much-needed technology advances. The resulting collaboration plays a very important role in finding solutions to the bottlenecks that currently prevent CPS from having greater impact on society; such solutions would also promote market uptake, open new markets and optimize the use of resources in the various industry sectors.

These communities have many cross-cutting challenges for future CPS. Embedded computing will evolve significantly and plays an essential enabling role for functional properties. For instance, the need to use specific sensors on a CPS and to timely process the relevant raw data onboard will need increased computational power. However, energy limitations introduce other constraints; only a holistic vision of CPS can help driving research initiatives. Moreover, the envisioned combination of 5G, beyond 5G and Smart Networks and Services/6G technologies with distributed and high-performance computing will pave the way towards a deep integration of future CPS in the computing continuum, where there are also direct links with enabling technologies, discussed presently, such as IoT and Systems of System. In relation to decentralization and decomposability, with distributed intelligence and emergent properties, an example research context would aim to solve/work on delays in physical, computing and actuation timing. This requires model design and simulation approaches to capture the full heterogeneity of the system and its contributing communities. Physical interaction with people requires a system to have high fidelity knowledge of its environment and its physical dynamics. This requires the technologies of the functional properties community, which in turn need integration with the safety and security measures set by the system-engineering community. In the future, achieving greater energy efficiency will pose an additional challenge for CPS. The evolution of CPS must consider their overall energy footprint to minimize their environmental impact. To achieve this goal, all aspects, components, and technologies related to CPS need to be carefully considered. This effort will involve traceability among the contributing communities. It is therefore evident that the

best way to advance future CPS is to further support integration and aggregation approaches for community collaboration.

Systems-engineering communities

The development of CPS requires a holistic development approach that brings together a wide range of disciplines. This includes the typical systems engineering disciplines, such as requirements engineering, architectural design, implementation and quality assurance including system-wide responsiveness, safety and security. The disciplines of this community are important in terms of both the CPS in general and individual systems engineering sub-processes, such as mechanical engineering, control theory, electrical engineering and software engineering.

In almost all of our application-driven future scenarios, like in autonomous driving and Industry 4.0, CPS must be able to fulfil their purpose to a large extent without intervention of human users [9]. According to the Society of Automotive Engineers (SAE) taxonomy for autonomous driving, we refer to such systems as highly automated or fully automated CPS [9]. Already today and even more so in the future, systems engineering is one of the core competence fields for building such highly automated or fully automated CPS.

In the case of highly automated CPS, it is necessary to have a more comprehensive understanding of the term ‘functional safety’. In contrast to the understanding of the term by the ISO 26262 standard, which essentially considers the malfunction of system components, highly automated CPS require an analysis of the interaction of a) the functionality of the CPS under consideration with b) its context (e.g. other CPS in collaboration). This analysis serves to detect possible safety threats resulting from the interaction between system functions and contextual conditions, such as the interaction between the autonomous driving function of a vehicle and the failure of the signalling system at an automated road intersection. This new understanding of functional analysis, which goes far beyond the requirements of ISO 26262, is the subject of the SOTIF standard [10].

It is important to understand that security threats can also arise from inadequate or non-compliant cyber security. Relevant cyber security standards, such as ISO 21434 [37] in the automotive sector, have recently attempted to take this into account. Corresponding measures to mitigate such security

issues then refer to the establishment of appropriate measures and technologies to increase the cyber security of the CPS to an adequate level. It becomes evident that systems engineering research in the CPS field needs to take holistic, tightly integrated approaches for safety and cyber security engineering that consider both the engineering of the CPS and the management of its operation.

These threats to safety must be identified during the development process and mitigated, e.g. by specifying suitable requirements or safety devices (safety monitors) which bring the CPS to a safe state should CPS fail to behave according to requirements or expectations. Since CPS often monitor and control technical or physical processes, control theory is a discipline of great importance in the development of such systems. In this context, the concepts of monitoring and controlling technical/physical processes are reflected in various artefacts of systems engineering. For instance, the requirements originated from the way the processes should be controlled, as well as from decisions made about the design of the necessary sensors and actuators or even about the design of the algorithm for the computational processes of the feedback system.

In order to be able to develop such complex technical systems consisting of software and hardware, seamless systems engineering processes are required, establishing techniques, methods and tools for challenges such as the following examples. Since CPS in many fields of application work together in dynamically formed networks at runtime to pursue higher-level goals, possible collaboration structures must be identified and analysed in requirements engineering. For example, in the development of autonomous vehicles, the collaboration structures in which these vehicles must operate should be taken into account. Examples of such structures might be vehicle convoys to optimize the flow of traffic or at automated intersections to ensure safe crossing of the intersection, even with high traffic volumes and in complex traffic situations. In collaborative CPS, the issue of coordinated decentralized monitoring and control of technical/physical processes is added; an example of this is the coordinated acceleration or deceleration of the various vehicles within a convoy of vehicles.

In the case of highly automated systems, the involvement of the human user is required in (a few) defined situations to ensure that the system is able

to fulfil its purpose of ensuring safe operation. The integration of the human user must be effective, i.e. the user interface of these systems must be designed in such a way that the human user is able to perform the necessary tasks according to the intention, as free from errors as possible and within the existing time restrictions. One might think here of the example of autonomous road traffic, where highly automated systems require the driver to take control of the vehicle when a critical driving situation occurs.

Enabling-technology communities

Internet of Things (IoT)

The Internet of Things community developed around the goal of providing a means for all devices to be globally connected via the internet. The name ‘Internet of Things’ was used in 1999 by Kevin Ashton during a presentation to his higher management at Procter & Gamble. He described IoT as a technology that connected several devices with the help of RFID tags (radio frequency identification) for supply-chain management [11]. In 2008 the first international conference on IoT took place in Switzerland, discussing RFID, short-range wireless communications, and sensor networks; today, these topics continue to represent the major technological research domain for advancing the IoT, gathering information about the real world that can then be made useful in some way [12].

Since 2010 it has been normal for many different devices to be in our homes to be connected to the internet. Connected devices are used extensively in the consumer domain. In 2015, to support advancement of IoT for industry, the European Commission created the Alliance for Internet of Things Innovation (AIOTI). Applying IoT to the industrial environment has been termed industrial IoT, or IIoT, and has the goal of optimizing production value while considering the many additional challenges related to safety, security and performance. IIoT technologies support interconnectivity with the internet in the context of these challenges, enabling not only networked smart objects and information technologies but also “optional cloud or edge computing platforms, which enable real-time, intelligent, and autonomous access, collection, analysis, communications, and exchange of process, product and/or service information, within the industrial environment” [13]. IoT technologies, in particular those for the IIoT,

will be standard constituent elements of future safety-critical frameworks.

Enabling the infrastructure to support distributed intelligence and information exchange is at the core of IoT, so supporting cross-community work on CPS decentralization, decomposability and human interaction is important. These are already areas receiving some focus from the IoT community [14], [15], as indeed is the case for bringing communities around an embedded computing backbone, with work considering edge-cloud computing [16] exchanges. As an enabling technology, IoT responds to support other domains which means its focuses change based on the latest domain challenges, corroborated in recent IoT road mapping activities that its landscape is changeable in nature [17].

Artificial intelligence (AI)

Autonomy will bring incredible new benefits to CPS, but there are major challenges that must be overcome. The intelligence that can be applied is limited by current approaches to certification, legal frameworks and (lack of) trust for such systems. There is also a fundamental mismatch between the approach to functional safety software (top-down, correct by design) and current approaches to deep-learning programs (bottom-up, data driven), as addressed by projects such as SAFEXPLAIN [18]. These need to be addressed while maintaining and increasing the safety of such systems (which calls for improved traceability of the influences between the contributing communities to CPS). Safety of such systems is a serious challenge as the levels of reliability achievable by ML/AI are simply inadequate for high integrity systems (safety integrity levels 3 and 4) by and are an active area of research attracting investment from public and private sector. Cyber security of such systems is even greater challenge as vulnerabilities of ML/AI are significant and widely spread.

Reducing or mitigating these limiting factors will be an enabler for many advanced AI technologies related to decision making, learning etc, for the operation of the systems. In parallel, the other communities can provide more robust technologies for systems that are evolving as a result of AI. Of course, there are identified routes for AI to become “more trustable”; these include explainability of actions in human language, and the application of AI to non-safety-related aspects of CPS like decision support for system design. A serious

concern is that the approach to making AI/ML adequate for use in safety-critical context seems to lack the “system context”. Judging whether an AI/ML-based component is adequate for use in a safety-critical system should be done for the system that it is intended to be used. To illustrate take the so called *safety monitors* as an example, devices meant to mitigate the risks from insufficient reliability in autonomous vehicles including of components built with AI/ML. In a recent paper Terrosi et. al. demonstrated that the accuracy of safety monitors is not merely a property of the monitor, but is significantly affected by the system it monitors (e.g., an autonomous vehicle).

A significant characteristic of CPS will be coordinated collaboration. This relates to the way components of a CPS coordinate with each other or with people for outcomes only achievable through such cooperation. AI can bring strong support here such as through the field of decentralized intelligence called multi-agent systems (MAS) [19]. Regarding design, the needs of CPS include the explicit representation of the environment and the need to represent abstraction layers, from the physical layer to the components and system, as CPS are closely coupled to the hardware elements of the system. Finally, it may also be necessary to represent the non-functional requirements, such as safety or resilience. Some MAS design tools, such as Tropos [20], if correctly used, may help to meet these requirements.

In terms of decentralized intelligence for CPS, there are many challenges to that need to be addressed, in particular methods for executing coordination. The whole system needs to be able to react in real time, which is not the case for most decentralized AI coordination protocols, which rely on negotiation, usually with no defined deadline for decisions [20]. As another example, finding ways to work with the functional property community on communication middleware for intelligent collaboration is likely another issue to need to be tackled.

High-Performance Computing (HPC)

High-performance Computing (HPC) consists of the aggregation of powerful computing resources for solving problems that require large computing power [21]. Recently, HPC technologies were only required in the context of traditional massively parallel “number crunching” applications like weather prediction, computational chemistry, or computational fluid dynamics. However, the latest

developments in low-power computing technologies [22] – required in the HPC industry to scale performance levels further – has facilitated the adoption of HPC technologies in a wide range of CPS applications.

Existing HPC platforms offer the computation capabilities needed by the most demanding CPS applications within an affordable power budget in domains such as automotive, space, avionics, robotics and factory automation. Centralized domain architectures that replace the traditional federated computing architectures – like those required by economically affordable autonomous driving systems – are only possible when HPC technologies are deployed. Single-chip high-performance embedded computing platforms reduce the traffic flow through CPS’ electronic networks and enable high-speed communication as required for processing vast amounts of information in real time. So this community will be important for consolidating the embedded computing backbone.

Furthermore, these technologies involve parallel processing, that is, splitting the tasks up into parts for several computers (or multiple cores) to process, thus reducing the time taken to complete tasks. This characteristic thus holds a direct relation with the CPS challenges of decomposability and decentralization – how tasks can be split up while ensuring safety and security for people, the system and its environment.

Unfortunately, the deployment of HPC in a CPS increases the complexity of the resulting system and may have non-negligible impact on the verification and validation costs of relevant system properties (e.g. safety and security). Thus, an effective exploitation of HPC technologies in cyber-physical applications requires at least either the development of new methodologies to verify and validate such complex systems or the adaptation of key technologies to the specific context, as explored in the EU-funded PROXIMA [23] and MASTECS [24] projects, for example.

Big Data

Cyber-physical systems are being driven by the combination of embedded and internet technologies and a vision of “smart anything everywhere” [25]. The blend of this cyber, physical (and social) data can help us to understand incidents and changes in our adjacent environments better,

monitor and control buildings and urban infrastructure, and provide better healthcare and care services for older people, among many other applications. To make effective use of the physical-cyber-social data, integration and processing of data from a variety of heterogeneous sources is necessary. A key objective for big data in CPS is to analyse very large, fast, and heterogeneous data streams, mostly from industrial rather than consumer environments. This can be achieved through machine learning, which is the most common technique used to extract information from the data.

The core Big Data applications in CPS are in varied fields, including energy utilization, city management, transportation systems and disaster management. For example, a smart transportation system would generate big data consisting of drivers' behaviour, commuter information, vehicle locations, traffic-signal management, accident reporting, automatic fare calculations, and so on. Robot-aided surgical systems (i.e. human-in-the-loop CPS) comprise a teleoperation console operated by a surgeon, an embedded system hosting the control of the automated robot, and the physical robotic actuators and sensors. Big Data methods can be used here for the modelling of surgical skills, for the detection and classification of surgical motions for automation and environment, and for the integration of this knowledge into control and automation of surgical robots.

In the operation of complex systems (e.g. aircraft and industrial processes), fault-detection and fault-isolation schemes are designed to detect the onset of adverse events. Such systems use big data methods (such as machine-learning classifiers) to enhance the diagnostic accuracy of the online reasoner on board the aircraft. Moreover, big data can be utilised in command and control with cyber-physical infrastructures for emergency services and defence.

The value of the Big Data community as a contributor to CPS products can only grow in the future due to increasing interest in data as an important business asset. The combination of heterogeneous data from numerous sources will require new applications for integration, query and analysis, along with embedded computing, high-performance computing, and data-reduction techniques. This remains an open research issue for CPS. The variety of types and sources of data will give rise to new kinds of data stores to sustain flexible data models.

Another important issue is that of remote storage of big data. Until now, cloud-based models have facilitated the storage and processing of big data sets, providing data accessibility and better IT power. However, this creates a centralized data store that does not scale in the CPS setting. To facilitate decentralized data storage and processing, a number of problems (e.g. replication, parallelism and requirements) arise. There is an urgent need for new approaches and techniques.

System of systems

The "System of Systems" (SoS) concept has been around for at least fifty years, but in the last twenty it has been an area of major concern. Following the description of its characteristics by Maier [26]; it is defined in ISO15228 as: "SoS...brings together a set of systems for a task that none of the systems can accomplish on its own. Each constituent system keeps its own management, goals, and resources while coordinating within the SoS and adapting to meet SoS goals" [27]. As for CPS, SoS represents a type of application, which can be the same, e.g. railway systems, as well as a technology domain - where the focuses are different.

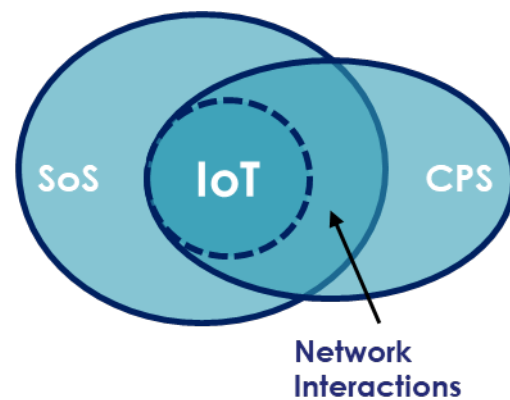


Figure 2: Technology relations of SoS, IoT, and CPS [35]

Broadly, one can consider SoS applications as independent systems that interoperate (work together) to achieve a purpose, with a significant amount of ubiquitous networking. In the case where they have extensive software control between safety-critical systems, the application itself is both a SoS and a CPS because they share common characteristics. Figure 2 describes the relationship between SoS, CPS, and the Internet of

Things. Where infrastructure interactions are supported by internet protocol, then the CPS is also described as IoT, which is necessarily always a SoS. There are also interesting SoS-CPS applications that interact through means other than the internet protocol (e.g. mechanical or electromagnetic interactions) and the engineer may need to guard against such interactions for safety or performance reasons.

However, from the technology perspective, CPS application research considers how all technology communities are integrated to create a system and its interactions, with the SoS technology community contributing to the coordinated collaboration aspect. This is a key property for future CPS, meaning that SoS research is indispensable for creating future CPS. In relation to embedded computing, the importance of localized processing, while maintaining a connection to centralized processing capacity, is recognized as a priority in areas such as edge computing, which uses SoS technology. This also links directly with the challenge of decentralization or decomposability where systems work together. A smart city is an example of human interaction and SoS, for example; it manages busy traffic at city junctions to minimize delays for drivers and pedestrians.

In 2012, INCOSE conducted a survey to identify “pain points” for SoS practitioners, i.e., the problems that kept systems engineers and managers awake at night [28]. The study indicated seven main areas of concern: SoS authorities; leadership; constituent systems; capabilities and requirements; autonomy, interdependencies and emergence; testing, validation and learning; and SoS principles. It is no coincidence that creating CPS includes these pain points, because they are concerned with networked, intelligent systems of high complexity. This suggests that the communities of SoS and CPS have areas of common interest suitable for collaboration.

Digital Twins and the Metaverse

Human-CPS interaction will also advance with the advent of digital-twin and “metaverse” technologies [29], in particular when CPS operate in close proximity or hand-in-hand with human operators. The metaverse will provide haptic feedback over robots that complement and advance human capabilities [30]. Human operators will receive visual guidance in their view of augmented reality, and will obtain the ability to project themselves into

the CPS they control. They will sense, act and interact through the impersonated system with other humans and with the environment in which the CPS operates. They will receive extended cognition and operating capabilities over swarms and manage the complexity of CPS hierarchies with ample application areas. Human caretakers may intervene in case of emergency or when service robots hit the boundaries of autonomy. At the same time, the fact that the environments surrounding a CPS are very diverse and unpredictable, will require that they also be incorporated in these virtual representations. Any kind of CPS autonomy risking damage to people or goods in the surroundings must be accounted for from a trustworthiness perspective.

Swarms will act in harsh environments on Earth, in space and on remote celestial bodies instead of exposing humans to the risks they have to take today. Examples include mining, nuclear-waste handling and reactor deconstruction, but also asteroid mining and exploration. Replacing the internet with a network of immersive virtual worlds, cyber-physical systems will allow the metaverse to bridge into reality, with all the benefits, but also all privacy, safety and security risks this entails.

Digital twinning is one of the enabling technologies for exercising such advanced control from the digital realm over the real, physical world. Digital twins are virtual models of reality that are continually updated about the actual state of their physical counterparts and which can enable decision-making that, in turn, leads to changes in the real world. The long-term goal of digital twins is to be able to capture the intentions and objectives of the physical twin, but also to improve overall performance through digital simulation, testing and monitoring how the real-world physical system will act in its environment. While the aim is to advance into a better future, this can threaten safety and security when not handled with utmost care. Thus, it will be inevitable for the metaverse and digital-twin communities to join forces with the CPS community to achieve real-time safe, secure, and cyber-attack resilient automation from the moment metaverse-enlightened CPS are designed and throughout their lifetime.

Production-environment influencer communities

Members of the production-environment communities are responsible for the industrial product

process and lifecycle. This includes enterprise policy and processes, decisions about technology usage and the evolving physical plant [31]. They drive the large-scale production of goods using equipment in the form of modular automated product lines. Such equipment typically combines mechanical, electrical, and software components; it also requires substantial initial investment and maintenance costs. Throughout its long lifecycle (15-30 years) [31], the equipment operator and component suppliers cooperate to repair and repurpose/upgrade parts at a minimal cost. This imposes several constraints on component models and their versions, which in turn constrains policy and process management.

In addition, the arrival of digitalization and the CPS revolution brings the “servitization in manufacturing” opportunity, a paradigm shift where manufacturers shift to offer product-related services, beyond just selling a tangible asset. In the above example of automated product lines, component providers could offer monitoring, online maintenance, repair, and overhaul services [32] among other value-added services. Service contracts generate more steady revenue compared to the cyclical product business, but, in general, organizations in manufacturing struggle to drive servitization [32], because the introduction of the new services incurs higher costs without proportional returns.

The adoption of digitalization tools and solutions and the development of innovative services leveraging the full potential of CPS require incentives and coordinated efforts among different partners. Research projects, partnerships in which early movers and less-digital companies cooperate to embrace servitization and adopt CPS tools, provide a nurturing environment, where decision-makers find that the “test-before-invest” concept is an incentive that helps lower barriers and can evaluate potential benefits. For example, in the H2020 HUBCAP [33] project, less digitally focused SMEs were able to pair up with model-based design providers to adopt digital innovation and enhance their solutions using model-based design technology.

Among the success stories, there is the example of the partnership between Mototok International GmbH, a provider of innovative aircraft tug solutions, and Evitado Technologies GmbH, a provider of LiDAR-based algorithms adding advances from the self-driving car industry to an already innovative CPS product. Other examples

show how advances were made in training for industry 4.0, the development of innovative organ preservation devices in the medical domain, smart textiles, and precision agriculture.

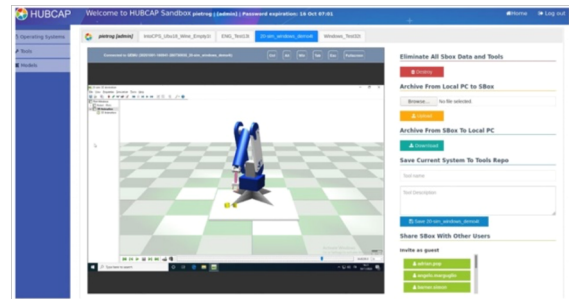


Figure 3 : Snapshot from the Sandbox showing SME asset

The prime innovative aspect of HUBCAP is a web-based collaboration platform that facilitates stakeholders’ access to computing resources and advanced CPS design and engineering solutions, by providing a cloud-based sandbox solution. The sandbox provides pre-installed models and tools, allowing companies to experiment with new tools and assets in a ready-to-use virtual machine available via a regular web browser, with emphasis on performance and interaction between partners. This is taken forward and combined with DevOps capabilities, also in a digital twin as a service (DTaaS) setting [38].

Production-environment community members are deeply involved with the cross-community challenges identified. There is a historical synergy with the development and advancement of embedded computing, which will continue in the future. This community is always demanding advancements in embedded computing, and advances in manufacturing also affect how we produce the embedded platforms of the future. Regarding decentralization and decomposability, there are several lessons learned and case studies in which cooperation and adaptation to local and greener processes promote research, discussion, and changes to manufacturing. Finally, this community has a particular interest in the challenge of physical collaboration with people. This interest is from both an internal perspective, covering topics such as human-machine interaction and collaborative robots, and an external perspective, where the potential for improvement from product usage data needs to be fully explored.

Market-influencer communities (society needs, regulation, standards, policy)

CPS are believed to have an enormous impact on many aspects of socio-economic life. Therefore, a number of stakeholders grouped here under the generic name of ‘market influencers’ will have a stake in shaping the future of CPS and of the contributing communities.

Societal needs may be described basically by means of individuals or groups putting forward requirements and benefitting from CPS. The individual appears here as the consumer who is, in one way or another, making use of either a product incorporating CPS, or elements of larger CPS implementations, addressing communities of end users in terms of mobility, personal life (general wellbeing), healthcare, leisure, environment, etc. Other needs may be identified in the area of public services offered at local and national government level, including education, healthcare services, community services, and operation of public institutions.

As well as responding to societal needs, however, CPS also pose new challenges. Some specific fields include education and employment, as CPS induce the obsolescence of certain professions and create new ones. Therefore education, including training and retraining will be affected, as will the employability of the existing and future workforce, which will have implications for the labour market and social security.

Regulation – both hard and soft legislation - will have to be adapted in order to govern CPS so as to ensure their smooth integration into society. However, given the rapid cross-border spread of CPS technology, international agreements might be needed, too, particularly if we consider the globalised nature of today’s value chains. Regulation will have to address the interplay between CPS actors (producers, consumers) as well the foreseen and unforeseen effects of the technology. Regulation is also supposed to be structured according to the societal needs that the technology is supposed to fulfil. A particular aspect of related regulation might address the human individual, chiefly in relation to human-machine interaction, which is anticipated to increase significantly in the coming years (intruding into both privacy and healthcare). The “must be implemented” regulation should be supplemented with recommendation-type measures of indicative nature.

Standards ensure interoperability and compatibility of products from different producers and allow the market presence of a large number of actors. Moreover, standards are important in order to set and describe safety levels and quality frameworks. To some extent, standards provide the technical base for legislation governing the area and also give room to innovation as usually standard specifications can be fulfilled in a variety of competing ways.

Policy aims to achieve certain results in a given field by reflecting society’s needs or goals. Public policy in particular is directed towards supporting certain areas through frameworks of development in terms of tax incentives, grants or even regulation. Policy also includes public investment in facilities or processes of general interest. A further aspect for consideration is policies aiming to increase employment in a differential manner within the given population (i.e. in favour of disadvantaged groups), or to ensure development of regions lagging behind. Such policies also set out to address issues of general interest like climate change (that can only be done at international level) or the environment.

Beyond public policy, one should take into consideration policies of generically named “groups of interest”. Pressure groups such as non-governmental organizations (NGOs) consumer associations also have policies for their vision and procedures to supporting their realisation, which can indirectly influence the market.

These “market influencer” stakeholders between them represent the conditions under which all the other communities operate for producing future CPS. The relevance of their involvement should be apparent, especially when considering the aggregative effects of contributing and cross-community technologies. Deficits in education in one community can have a knock-on effect on other communities. Training approaches and certification can be a deciding factor in the sustainability of mixed-community technologies. Policy can evolve approaches and perspectives that enhance behaviours supporting longer-term governance or culture, providing resilience, value generation and trust in new technologies.

Research orchestration for large-scale safety-critical systems

With respect to coordinating CPS research as an application domain, additional approaches and orchestration should be introduced. This is because the application-domain perspective is based on the product side, with cumulative effects being considered through the aggregation of layered contributions from the stakeholder communities. Another issue is that disruptive discoveries, technologies or developments might influence the cycle of research. For example, if significant progress is made on quantum computing, or discoveries in material/biological science, that could make sensors more different.

Orchestration of research is particularly about knowledge management, longer development cycles, persistence and refinement of multi-disciplinary approaches for collaboration between communities. Take the example of constructing a building where a new team takes over every few months. Limited progress can be made without guidance at a higher level. This is similar for advancing CPS research. Persistence of acquired interaction techniques, between project collaborations, is significantly more difficult to maintain. For instance, usability and sensor experts have specific languages for their domains.

Therefore, approaches that support collaborations and which have been developed during collaborations should be taken, refined, and applied in subsequent collaborations of different groups. A dedicated CPS research instrument could advance this concept, in conjunction with future CPS support action projects. Projects themselves will also need to provide environments with favourable conditions for aggregative research considering the multi-dimensional challenges, with conditions significantly different to those for developing component technologies.

Considerations for future CPS projects

For advancing CPS research as a technology domain, useful mechanisms already exist. For example, there have been projects following the standard approach, which gathers technology providers around one or more CPS-related use case. If awarded funding, the partners then work together for a few years to bring their technologies closer to market deployment (i.e. advancing “technology readiness levels” or TRL).

Cascade funding, where funded projects themselves fund smaller initiatives, has also shown itself to be a useful means for transferring component technologies for CPS, because the smaller initiatives are directly managed by companies looking for particular solutions.

However, for the application domain side of CPS research, new project approaches and higher support mechanisms also need to be introduced, enabling the multi-dimensional challenges previously discussed to be tackled. The characteristics that are believed to be essential in such projects are:

- Use cases: physically interactive and collaborative systems; of relevance to all communities, likely to be uniquely large industry or with integrated small-medium enterprises. Supplied also with the intention of advancing industry-side “industrial readiness levels” of production and product lifecycles for new technologies.
- CPS centre of gravity: all projects addressing the multi-dimensional challenges between communities should interface on work advancing real-time safe and secure automation, including interrelations, for CPS design and operation.
- Cross-community challenges: projects on application domain research should focus on grand challenges that need contributions from each community. Proposed call topics include:
 1. Embedded computing backbone
 2. Decentralization and decomposability
 3. Physical collaborations with people
- Developing the support environment: tools and approaches are required not only by industry, but also by researchers to support engagement of the different CPS stakeholders and perspectives. We propose that such projects include some dedicated work (a work package) that develops support for collaboration on the multi-dimensional challenges.
- New approaches established iteratively: orchestration approaches should be implemented in a manner that can be refined. We should avoid ‘one-hit wonders’ that seek to solve everything at once. A second iteration of such projects could also include smaller spin-offs and initial stage smart city investigatory projects.

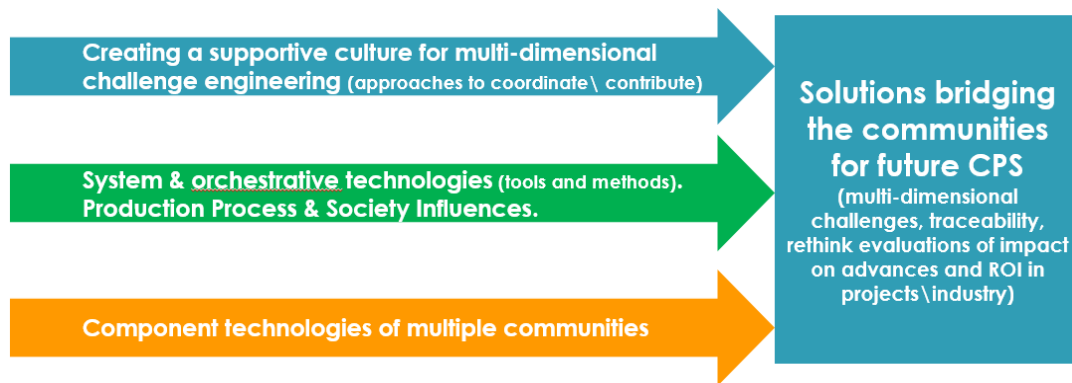


Figure 4: Stakeholder inputs to projects.

Contributions of the communities to be these projects can be visualised, as shown in

Figure 4, to involve the technology component providers, the influencers/aggregative technology providers and those developing the culture and support environment. This provides the means to advance the CPS aggregation techniques which are required to address the multi-dimensional CPS challenges.

These HiPEAC-proposed project characteristics (for application-domain research) relate directly to previous CPS community recommendations to the European Commission, including trustworthy and societal scale CPS, ethics data protection and liability, CPS engineering, interoperability, complexity, edge computing, humans-in-the-loop, co-engineering of system properties and enhancing uptake of CPS technologies.

Considerations for future advisory coordination and support actions for the CPS communities

The European Commission funds coordination and support actions (CSAs) to accompany, coordinate and stimulate innovation in particular technology fields and their communities. A particular challenge for a CPS CSA is that it is in fact a multi-community subject. This is because, as discussed, while CPS is a technology domain with specific complex challenges related to cyber and physical integration and cyber to physical plan realisation, CPS is foremost an application domain. This is of consequence because CPS and other technologies can be much more difficult to apply to the final

systems without also advancing the means for their combination.

To support application domain research projects, future CPS CSAs will likely support the transfer and synchronization of project environments, support the “big picture” metrics of aggregations in CPS and specific return-on-investment (ROI) valuation techniques to pre-empt industry needs. In particular, they will support a focal point for all the contributor technology and influencer communities.

Research Instrument for technology orchestration: Supporting projects, funding programmes and industry

A team providing support across relevant research programmes, dedicated to supporting the channelling of different community contributions into safety-critical applications, in particular, would be very beneficial for supporting in particular technology orchestration/aggregation and technology uptake by safety-critical systems/CPS in general. This can only really be properly realized if this team provides an ever-present pivot for CPS projects and CSAs, developing the support environment required to manage the multi-dimensional challenges.

They would have two support roles: the development and investigation of concepts that are provided by projects and programmes that had been identified as useful support assets (to the projects), but their implementation being normally unattainable in the scope of the projects and programmes.

On the development side, support to programmes would include, as an example, enhanced tool techniques for directed communication (the right information, to the right people, at the right time – especially for start-ups). The research instrument team provides prototyping tools to relevant CSAs that would support deployment in technology orchestration projects, who then test and further develop the tools. Support for the creation and testing of tools largely depends on results from the investigatory side. Some examples include:

- Inter-community supports like wiki-type project glossaries to manage the multiple perspectives (e.g. mediation between safety/security, medical/railway, SMEs/large enterprises).
- Multi-community access like digital passports, allowing users to access and test many research tools with the same account.
- Improved techniques like supporting management of intellectual property rights.
- Connecting contributions, such as a holistic view of open-source tool advancement across projects.

The investigatory side considers and proposes enhancements from the product-side perspective, for projects, programmes and industrial policy. These would be potential assets for promoting in particular aggregative technology uptake and longer-term profitability. Investigations would consider enhancements outside our normal fields of operation. Potential concepts include:

- Supporting the project environment for capitalization on and continuation of knowledge from multi-stakeholder interactions. Approaches for iterative improvement. Incentives, performance measures, mentoring.
- How are CPS-specific and aggregative technologies advancing, what is the funding flow to the contributing communities? Studies on benefits – but also consequences of lack of funding.
- Managed contributions, e.g. open-source results – rather than a default expectation, should be with respect to conditions (such as business model, maintenance, community building).
- Considerations for adapting the destination (industrial processes) to the new technologies; how to lift constraints at the product-side.

- Lighthouse initiatives within programmes (advancing structuring and management policies) may provide ideas to be explored.
- Currently, technology readiness levels (TRLs) measure the advancement of individual components rather than aggregations of components. A complementary approach, let us say aggregative-TRLs, is therefore required. This is not to be confused with the “integration readiness levels” measuring the interface between technologies (how they connect), rather than aggregating technologies (managing their combined effect).
- Supporting the development of a body of knowledge and teach the science of CPS engineering.
- Balancing local/national/European interests across networks. For instance, cross-border Digital Innovation Hubs (DIH) could complement the specific interests of regional or national DIHs.
- Policy on protection of EU business data (~B2B GDPR). CPS representation would be relevant here to consider the effects of such a policy on CPS technology advancement.
- Studies to advise/encourage industry towards longer-term strategies. This may also include changes in government regulation to shift from short-term competition of yearly quotas towards longer-term and more profitable competition and managing incentives where average employee turnaround is 3-4 years. No CPS-specific studies on corporate evolution seem to exist yet.

The proposed ways forward through this higher-level support from a CSA and a research instrument not only enables significant advancements for future large-scale safety-critical systems research, but also addresses the recommendations made by previous projects for CPS technology (such as Platforms4CPS [34], which represents an update of several roadmaps). These earlier recommendations included: collaboration and defragmentation of siloes; public understanding of the importance of CPS; supervisory support to draw together a common body of knowledge; and developing talent in order to maintain Europe’s leadership and sovereignty of diverse technology aggregations for multi-domain applications including transport, manufacturing and health.

References

- [1] S. Paul, "D3.4.4. MERgE - Recommendations for Security and Safety Co-engineering v3 partA," 22 April 2016. [Online]. Available: <https://itea4.org/project/merge.html>. [Accessed 26 Nov. 2023].
- [2] "RITICS: Research Institute in Trustworthy Interconnected Cyber-Physical Systems," [Online]. Available: <https://ritics.org/>. [Accessed 26 Nov. 2023].
- [3] "ICRI_SAVe," Intel, [Online]. Available: <https://www.icri-cars.org/>. [Accessed 26 Nov. 2023].
- [4] "IET Code of Practice: Cyber Security and Safety," Institution of Engineering and Technology, [Online]. Available: <https://electrical.theiet.org/guidance-codes-of-practice/publications-by-category/cyber-security/code-of-practice-cyber-security-and-safety/>. [Accessed 26 Nov. 2023].
- [5] "MERgE: Multi-Concerns Interactions System Engineering," ITEA 4, [Online]. Available: <https://itea4.org/project/merge.html>. [Accessed 26 Nov. 2023].
- [6] "SESAMO: Security and Safety Modelling on CORDIS," European Commission, [Online]. Available: <https://cordis.europa.eu/project/id/295354>. [Accessed 26 Nov. 2023].
- [7] "AQUAS: Aggregated Quality Assurance for Systems," [Online]. Available: <https://aquas-project.eu/>. [Accessed 26 Nov. 2023].
- [8] "CPS Swarm," [Online]. Available: <https://www.cpswarm.eu/>. [Accessed 26 Nov. 2023].
- [9] SAE International, "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles J3016_202104," 30 04 2021. [Online]. Available: https://www.sae.org/standards/content/j3016_202104/ [Accessed 26 Nov. 2023].
- [10] International Organization for Standardization, "ISO 21448:2022 Road vehicles — Safety of the intended functionality," 2022. [Online]. Available: <https://www.iso.org/standard/77490.html>.
- [11] M. Roberti, "What Is the Internet of Things?," RFID Journal, 2016. [Online]. Available: <https://www.rfidjournal.com/question/what-is-the-internet-of-things>. [Accessed 26 Nov. 2023].
- [12] "Internet of Things 08: International Conference for Industry and Academia," Zurich, 26-28 March 2008. [Online]. Available: <https://iot-conference.org/iot2008/prg/program.pdf>.
- [13] H. Boyes et al, "The industrial internet of things (IIoT): An analysis framework," *Computers in Industry*, vol. 101, pp. 1-12, <https://www.sciencedirect.com/science/article/pii/S0166361517307285>, 2018.
- [14] K.H. Wöhnert et al., "Secure Cyber-Physical Object Identification in Industrial IoT-Systems," *Procedia Manufacturing*, vol. Volume 51, pp. 1221-1228, <https://doi.org/10.1016/j.promfg.2020.10.171>, 2020.
- [15] S. Sachdev et al., "Voice-Controlled Autonomous Vehicle Using IoT," *Procedia Computer Science*, vol. Volume 160, pp. 712-717, <https://doi.org/10.1016/j.procs.2019.11.022>, 2019.
- [16] A.J. Ferrer et al., "Admission Control for Ad-hoc Edge Cloud," *Future Generation Computer Systems*, pp. 548-562, <https://doi.org/10.1016/j.future.2020.08.024>, 2021.
- [17] E.M. Jakobsen et al, "NGIoT D4.6: Report on IoT business model innovation patterns and acceleration support activities," 30 September 2022. [Online]. Available: https://www.ngiot.eu/wp-content/uploads/sites/73/2022/09/EU-IoT_D4.6-IoT-BM-patterns-and-acceleration-support.pdf. [Accessed 26 Nov. 2023].
- [18] "SAFEXPLAIN: Safe and explainable critical embedded systems based on AI," CORDIS, European Commission, [Online]. Available: <https://cordis.europa.eu/project/id/101069595>. [Accessed 26 Nov. 2023].
- [19] M. Wooldridge, *An Introduction to MultiAgent Systems*, 2nd Edition, Hoboken, New Jersey: Wiley, May 2009.
- [20] C. Cares et al, "Agent-Oriented Engineering for Cyber-Physical Systems," *ICITS 2019: Information Technology and Systems in Advances in Intelligent Systems and Computing*, vol. 918, pp. https://doi.org/10.1007/978-3-030-11890-7_10.
- [21] "High Performance Computing," European Commission, [Online]. Available: <https://digital-strategy.ec.europa.eu/en/policies/high-performance-computing>. [Accessed 26 Nov. 2023].
- [22] N. Rajovic et al, "Supercomputing with commodity CPUs: are mobile SoCs ready for HPC?," in *SC13: International Conference for High Performance Computing, Networking, Storage and Analysis*, Denver, Colorado, 2013, <https://dl.acm.org/doi/10.1145/2503210.2503281>.
- [23] "Probabilistic real-time control of mixed-criticality multicore and manycore systems (PROXIMA)," [Online]. Available: <http://proxima-project.eu/>. [Accessed 26 Nov. 2023].
- [24] "MASTECs: Multicore Analysis Service and Tools for Embedded Critical Systems," [Online]. Available: <https://mastecs-project.eu/>. [Accessed 26 Nov. 2023].
- [25] European Commission, "Smart Anything Everywhere," [Online]. Available: <https://smartanythingeverywhere.eu/>. [Accessed 26 Nov. 2023].
- [26] M. W. Maier, "Architecting principles for systems-of-systems," *Systems Engineering*, vol. 1, no. 4, pp. 267-284, 1998.
- [27] "ISO/IEC/IEEE 15288:2015 Systems and software engineering — System life cycle processes," May 2015. [Online]. Available: <https://www.iso.org/standard/63711.html>. [Accessed 26 Nov. 2023].
- [28] J. Dahmann, "1.4.3 System of Systems Pain Points," in *INCOSE International Symposium*, Las Vegas, Nevada, 2014, <https://doi.org/10.1002/j.2334-5837.2014.tb03138.x>.
- [29] S-V Rehm et al, "The Metaverse as Mediator between Technology, Trends, and the Digital Transformation of Society and Business," *Journal of Virtual Worlds Research*, vol. 8, no. 2, pp. 1-6, <https://doi.org/10.4101/jvwr.v8i2.7149>, 2015.
- [30] Z. Sun et al, "Augmented tactile-perception and haptic-feedback rings as human-machine interfaces aiming for immersive interactions," *Nature Communications*, vol. 13, pp. 5224, <https://doi.org/10.1038/s41467-022-32745-8>, 2022.
- [31] S. Braun et al, "Requirements on Evolution Management of Product Lines in Automation Engineering," *IFAC Proceedings Volumes*, vol. 45, no. 2, pp. 340-345, <https://doi.org/10.3182/20120215-3-AT-3016.00060>, 2012.
- [32] M.M. Herterich, "The Impact of Cyber-physical Systems on Industrial Services in Manufacturing," in *7th Industrial Product-Service Systems Conference*, Saint-Étienne, 2015.
- [33] P.G. Larsen et al, "A Cloud-Based Collaboration Platform for Model-Based Design of Cyber-Physical Systems," in *SIMULTECH 20: 10th International Conference on Simulation and Modeling Methodologies, Technologies and Applications*, Online, 2020, <https://www.scitepress.org/Link.aspx?doi=10.5220/0009892802630270>.

- [34] "Platforms4CPS," [Online]. Available: <https://https://cordis.europa.eu/project/id/731599>. [Accessed 26 November 2023].
- [35] M. J. d. C. Henshaw, "Systems of Systems, Cyber-Physical Systems, The Internet-of-Things... Whatever Next?," *INSIGHT*, vol. 19, no. 3, pp. 51-54, 2016.
- [36] "AIDOaRt," [Online]. Available <https://www.aidoart.eu/>. [Accessed 3 November 2023].
- [37] International Organization for Standardization. ISO 21434 "Road vehicles – Cybersecurity engineering". Norm. 2021. <https://www.iso.org/standard/70918.html>
- [38] P Talasila, C Gomes, PH Mikkelsen, SG Arboleda, "Digital Twin as a Service (DTaaS): A Platform for Digital Twin Developers and Users", Preprint <https://arXiv:2305.07244>, 2023
- [39] C. Robinson. "Understanding cyber-physical systems among many communities: Large-scale safety-critical systems". In M. Duranton et al., editors, *HiPEAC Vision 2023*, pages 27-31, Jan 2023. <https://doi.org/10.5281/zenodo.7462025>
- [40] Robin E. Bloomfield & John Rushby. (2021). Assurance 2.0: A Manifesto. <https://arxiv.org/abs/2004.10474v3>

Charles Robinson is Research Projects Leader in critical embedded systems at Thales Research & Technology, France.

Rajendra Akerkar is Professor and Head of Big Data Technologies at Western Norway Research Institute.

Djamila Aouada, is Professor and Head of CVI2 Group, SnT Faculty, Université du Luxembourg.

Alessandra Bagnato is Research Scientist and Head of Modelio Research at Softeam (Docaposte Group).

Miklós Györffi is Senior European Affairs Analyst at the Hungarian Research Network and former staff member of the European Parliament.

Michael Henshaw is Professor and Programme Director in Systems Engineering, Associate Dean for Teaching, Loughborough University.

Peter Gorm Larsen is Professor and Head of the DIGIT Centre, Department of Electrical and Computer Engineering, Aarhus University.

Carles Hernandez Luz is Senior Researcher in Processor Designs for Safety-Critical Systems, Universitat Politècnica de València

Hugo Daniel Macedo is Researcher in the DIGIT Centre, Department of Electrical and Computer Engineering, Aarhus University.

Claudio Pastrone is Head of Connected Systems and Cybersecurity research domain in LINKS Foundation.

Peter Popov is Associate Dean (International), School of Mathematics, Computer Science and Engineering, City University London.

Claudio Sassanelli is Professor Department of Mechanics, Mathematics and Management, Politecnico di Bari.

Marcus Völp, is Professor and Head of CritiX Group, SnT Faculty, Université du Luxembourg.

Thorsten Weyer is a Professor of software engineering, Technical Hochschule Mittelhessen (THM), Giessen.



Internet giants are betting fortunes the size of Latvia's or Slovenia's GDP on Large Language AI Models for returns much bigger: no sovereign can afford to let corporate machinations dominate the evolution of humanity's social code.

Open-Source AI, why it is the best way forward for Europe

by Thomas Hoberg

Key insights

- The explosion of humanity was fueled by its ability to create 'social code' whose evolutionary pace and expansion operates vastly faster than genetic or animal behavioral codes: shifting even parts its evolution to computers may be as transformative as the invention of speech
- The degree and range diversity of our social code diversity is much bigger than any individual can imagine. The quest for efficiency through scale can exert unwanted and dangerous homogenization pressure
- Opening source and sharing knowledge is one of the most potent measures to release tension
- The vast majority of humanity's digitally accessible global knowledge base is coded in English, local sovereignty requires both making local language content accessible to others (push) and making global content accessible in the local language (pull)
- G42s Open-Source AI initiative shows how to progress both directions with Arabic
- Corporate AI giants grow out of control and need immediate countermeasures to retain sovereignty

Key recommendations

- Investigate G42 JAIS project, because it could be a base for a near global collaboration or blueprint for a similar EU initiative
- Collaborating with partners that are furthest away from a Western ethos, enables AIs to incorporate social code from the largest range of populations and then provide service to them, because it helps to eliminate the *invasiveness* of foreign social code
- Giant AI models in foreign corporate hands are 'invasive species' with potentially catastrophic consequences for local societies: they need to be countered via regulation prescribing ground-truth partitioning, open source, and interdiction of unauthorized data grabs

Evolving at the Speed of Code

Ever since Darwin had us not just come out of the trees, but with living cousins still left there, identifying what justifies our innate sense of human superiority became ever more involved.

Using or fabricating tools isn't even limited to our cousins, far more distant relation like raven manage that. Even our sense of self, awareness of us vs others, including empathy, which unfortunately always seems to require more effort than enmity, isn't at all unique to our kind. It is found with ever more species who also develop culture, which emerges as soon as this mix of awareness, emotions and knowledge come together and create motivation to pass it on to offspring as social code via some type of communication.

Its most recognizable manifestation is play, an instinct that has cubs running through simulations of a future where they need to take care of themselves and their charges and where they constantly hone, evolve, and seek to improve skills and the social code they inherit from parents and peers. That play involves demonstrative teachers, who double as immensely dedicated critics, giving very immediate, sometimes gentle, sometimes forceful feedback via anything from posture to pain: you don't need to take their *word* to get the message and work that into your knowledge base!

And while those abilities build on a giant genetic base, it's the much faster pace of social code evolution, which accelerated the emergence of higher life forms and finally allows us place a defining hurdle between us and our cousins: speech as a much more efficient means of communication than live demos.

While communication has been immeasurably important for humanity, we may not be able to claim exclusivity nor the lead. The jury is still out if whales or octopus can't in fact achieve much higher levels of information transfer: there are indications they might in fact have been "videoconferencing" long before we started "texting"¹.

True break-through discoveries, with the potential of completely transforming the future of an

individual or its group vs the rest of the planet have been made... again and again.

And then lost, as they never made it out of the individual's head nor beyond immediate kin, because they could not be communicated or preserved beyond immediate reach.

It took eons and millennia to progress language and abstractions to the point where tales allowed reliving an entire experience. The ability to abstract was urgently needed to manage ever greater bodies of knowledge, and it grew into another set of cousins, religion and science, which are at ever greater odds and entangled in politics.

These huge tomes of social code soon exceeded average individual capacity and led to sages, shamans and minstrels specializing in their upkeep and dissemination. The ability to preserve them with fidelity outside a mortal brain, first in writing, then in printing, and more recently in a form where they continue activity and evolution on a machine, has each brought orders of magnitude to how effectively social code can disperse and potentially reign supreme: the speed at which social code spreads is both a measurement and a cause for its success—which is always temporary, but tends to evolve most rapidly when scale provides room.

Code & Context

As it turns out, fidelity of recording over space and time can add as many issues as the precision lost in retelling, where critical truths—but also the more egregious nonsense, could fade by accidentally or conveniently leaving them out of the telling; we fail to recognize just how many assumptions from our every-day life we take for granted or perennially true. Those are then neither recorded nor put into context, creating deadening strife instead of valuable synergies when scriptures are interpreted far beyond the natural expiry date of the parts that got written down.

To illustrate, I once learned of a people with who time and space were merged into an ever-present spatiality that I found indelibly intriguing. For them the past was straight ahead for the evident reason that it was known and thus visible, while

¹ Whales and dolphins may broadcast abstracts of sonar imagery from an individual viewpoint towards a group to enable coordinated action. Octopus can essentially project video patterns on their skin

enabling an optical communications channel. The achievable symbol rates in both cases may be far in excess of anything possible verbally.

slowly blurring with distance. The future is unknowable and therefore obviously in the back where it cannot be seen. Both have a fixed point on the compass, which I don't remember and they can't imagine not to. It could have been the future going south or the past chasing the sunset, most important is that it's shared and identical amongst them all. It obviously does not mean that turning a corner steers the flow of time. Mostly it means that going [in the direction of] Forward or Backward is not an individual perspective, but an extremely strong shared sense of direction.

With direction and time fused so deeply into everyone's mind, they lose orientation as likely as we'd miss time turning backward: they do not get lost, even in the dark or a cave. As they move with full awareness through this space-time continuum on their feet, they know exactly how they are heading and which way to return.

Perfect orientation may seem a bonus, but imagine trying to converse with a member of that tribe, when it's not just about replacing words of one language with those of another, but where the two sides perceive the world in a completely different manner. They might come to regard us as Flatlanders [1] who can only see a circle where there is really a sphere, lacking awareness of a full dimension of the time-space that surrounds us. They might in fact have far less trouble conversing with whales and their sonar minds than we have.

Yet traveling the high seas, going into space, or just falling asleep in a night train may well be torture to them. Being plucked from solid ground and moved unconsciously across the time-space continuum would be like being robbed of your mind. Waking up without a perfectly valid spatial reference, would be as frightful as it is unfortunately for many people with Alzheimer's who struggle to remember their recent social context. We might fear such people unfit for the modern world, but their youngest generation might just judge us as utterly incapable of reaping the full potential and value of a spatial Internet².

It is very hard to step back and realize just how much of our social context, knowledge, prejudice, bias etc. we constantly carry with us and unwittingly assume and embed as we try to record or

write the most objective and scientific truths into code or content we create.

When the earliest electronic computers were used by the Allies to replace their human variants to more efficiently compute projectile trajectories to end WWII, not even the Axis disputed the numbers, only the goal. But today's IT isn't about the less contentious physics or accounting. With the rise of social networks, the vast majority of all computing capacity being dedicated to digital media content and as ever larger parts of that becoming artificially created, very little of it has a shred of objectivity left. And where von Neumann might have eliminated the theoretical distinction between code and data decades ago, machine learning models obliterate the distinction between content and code, also merging science and religion, if they were ever separate, as enlightened minds liked to believe. And as digitalization engulfs all but the most personal transactions in life, these models become governing bodies of social code, exert a force that becomes immensely political through sheer mass adoption.

Like all code, from genetic over social, cultural, scientific and software, these machine learning models are immensely expensive to create, but much cheaper to replicate [2]. So, once they offer a significant enough evolutionary advantage, they spread with the acceleration offered by the base technology, machine learning models beating sex, social code evolution and now even human programmers in generational speed. And looking back at how printing, machines, and programmable computers have transformed societies, it's very clear that transferring significant parts of social code to machines not just for storage and dissemination, but for *execution and evolution* will transform all the still very different societies on the planet.

One might argue that slavery was not abolished because after millennia owners suddenly developed a conscience, instead the industrial revolution enabled machines as a cheaper substitute. Manual work and even slavery isn't gone in all forms and places, neither will machines nor machine learning will replace humans everywhere. But judging by the huge transformational power already exerted by global content industry, where

² See 'The Spatial Web: Interconnecting people, places, things and AI for a smarter world' and [20]

Western ethos exerts much more influence than population distribution would indicate, it's clear that control and ownership over the social code embedded in these models from the content they digest is absolutely crucial for sovereigns of any scale and persuasion. And those sovereigns who aim for and can afford a longer-term perspective took note and turned to action.

China has a very long tradition of regarding scale as a power base and adapted with urgency when machines and industrialization allowed the far less populous West to unhinge what it believes the natural balance during the last century. Closely managing knowledge and its dissemination was deeply ingrained in its feudal history, and it seems perhaps even more important today, where scale is regarded as the most intrinsic factor to control. At the low end, anyone is allowed to have and even voice their own opinion. But as soon as you share it with five people or more, any but the officially sanctioned opinions are being filtered out. From the top China aims to employ its giant domestic market potential to the fullest, ensuring scale via a relentlessly homogeneous social code dictated by the current sovereign over a population base that still retains traits of a rather diverse linguistic, cultural, and ethnic past.

The Indian subcontinent may host an even greater diversity and continues to attempt standardizing its vast diversity of social codes into only two major dissenting variants with more vengeance than is peacefully absorbed. But both Asian giants, who represent the majority of humanity, as well as others in the East much nearer to Europe, seem to ignore that this diversity isn't just a historical burden but often evolved with a high degree of sophistication as a means of *conflict avoidance* when heterogeneous populations grew in a mixed environment.

Dense populations that survived for centuries often are not marked by the thorough homogeneity one might expect from a "pressure cooker", but via a highly evolved mesh of subcultures whose peaceful cohabitation is made possible by distinct food taboos, job differentiation, castes, religions, and feudal classes with very distinct bodies of social code that coexist highly differentiated for that reason. The only safe thing to say about them is

that they are highly political and thus cannot be universally applied³.

The European Union was founded after centuries of debilitating warfare as a means to enable peaceful conflict resolution in an area where feudal sovereignty was often completely decoupled geographically, linguistically and tribally from the subjects through generations of marital diplomacy.

While that intermarriage at the highest level of society imbued Europe with a rich legacy of shared culture which trickled from nobility via the bourgeoisie to the greater population, from the Middle Ages right through the enlightenment and industrialization, it did not close the tribal rifts below which expanded into nations. When industrialization created a productivity explosion that manifested in an often brutal expansion of these national social mores into colonies, that also brought the new nations into conflict over social code variations their highly interbred sovereigns [3] would hardly deign to squabble over—but it had their subjects march to war and death in millions.

While its former colonies inherited and still echo much of that cultural legacy over vast regions of the planet, Europe itself attempted a reboot after two terrible wars where religion, language, ethnicity, and most other matters that had been the source of past contentions were firmly put under lock: yes, they could all be regarded as rich cultural assets to be shared for the pleasure and benefit of all; no, none of them would be permitted to claim superiority or exclusivity at the cost of all others.

Mumbai's Dabbawala [4] food supply system may be hard to beat in terms of the diversity per surface area, but the highly modular **institutionalized** social code of Europe, which aims at sharing all elements where that seems to provide a benefit, while retaining locality where it reduces potential conflict, is still rather unique in this world.

US and Chinese internet giants have used their giant domestic markets to create awe inspiring head starts in AI, but once they try grow beyond and with scope of digitalization still expanding, the friction between cloud code and ground rules becomes too large, the lack of modularity and ability

³ high degrees of equalization pressure have led to revolutions, which then just reset into different strata

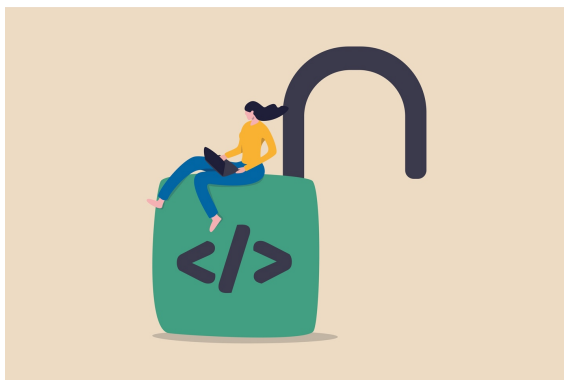
to support a diversity of social codes becomes an impediment.

What remains, is the giant cost of developing new code—now including social code, content and AI models—and the need for scale to spread the effort. And what needs to change is the hegemonic mindset of its creators: wherever humanity couldn't just dominate, it had to set down and negotiate for areas of collaboration vs exclusiveness of control.

And when it comes to planning ahead and for the greatest scale of collaboration achievable, it may be better to test such collaboration with partners where the social code gaps are much bigger than with those who we historically shared sovereigns or empires with. It needs to support a degree of collaboration with potential, even active enemies.

Open-Source Altruism vs Corporate Profit

Open Source at its heart is a decision to give up exclusive control over know-how or technology and share it instead with a wider group, often as part of a trade. Just like trade as an alternative to battle, it has existed for eons, it has built civilizations, but it has never stopped being weaponized, either.



At the time of this writing language models are being published as open source by companies that have invested enormous resources into their creation, not because sudden altruism compels them, but because they want to staunch a near exclusive flow of all attention and revenue to OpenAI-Microsoft as the technology leader, which would allow that company to out scale all potential

competition for a technology lead that would elevate it into an AI mainframe.

Very much like open-source Linux diverted financial and intellectual resources from the Unix mid-range companies, the aim is to avoid a stranglehold of exclusivity on a base technology where the current underdogs might have fallen behind. Open sourcing their 2nd rate assets reduces the leader's ability to monetize its lead and allows competition to continue at higher level race, where hopefully those losses are recouped.

Underselling the competition to eliminate it is probably older than Rockefeller's Standard Oil⁴, while the speed and disembodied nature of global software and content make it hard to combat when sovereigns are at odds. But the more exponential than linear cost of AI leadership make it difficult to sustain, which why the current top combatants fight all the harder for exclusivity, just as the open-source competition is all bent on keeping them from attaining a stranglehold.

Analysts (such as [5]), using what public information is available, find it unlikely that OpenAI and Microsoft have yet turned a profit for their AI ventures and are racing either for a GAI⁵ singularity or a practical monopoly to recoup their investments. While OpenAI's original goal as a non-profit corporation was proclaimed altruistic, Altman and his fellowship of employees and investors believe they deserve more of its valuation and are more inclined to raise the stakes than let competitors or regulating skeptics catch up, especially when Google may have an almost 'singularity' advantage in the form of their proprietary TPU series chips and infrastructure, that may scale both higher and far more efficiently than Nvidia as the otherwise undisputed champion. With Nvidia's top AI GPUs already apportioned mostly to Microsoft and very few other big players for the next years, that market has ceased to exist for lack of purchasable product. It's easy to see why Microsoft and Altman would want to exploit the position they dearly paid for already, and which may not last for long.

GAI capabilities in corporate hands, that find it difficult to self-regulate with competitors at their heels, deserve the slowdown the OpenAI skeptics

⁴ The main inspiration for US anti-trust law [21]

⁵ Also acronymized AGI, it stands for the threshold where an AI might exceed a qualified human in problem solving capability

attempted, just because it is likely to be as revolutionary as the steam machine. But that requires outside support from sovereigns. While corporate exclusive control over an emerging technology can lead to that technology never quite realizing its potential on its own, because consumers grow wary of feeding a giant that tends to become more beastly as its appetite grows, it becomes more difficult, when ubiquitous platforms like Microsoft's Windows and Office are infected with it before regulators react appropriately.

But then AIs, like the machines that started the industrial revolution, neither purchase products nor pay taxes for lack of income, while the resources they turn into emission may not be appropriately taxed. Without an economy to sustain them, GAI may see more short-term success as a vehicle of political power than be economically or socially sustainable.

Leveling a playing field where competition has ebbed away or is no longer generating taxable value is a strategy also often employed by governments, which might regulate or nationalize infrastructures like transportation, water or electricity grids to actually boost a more competitive economy that builds on top. And in the current AI race, one contender might inspire Europe.

Another Union's Quest for Sovereign AI

A century ago, the Arabian Peninsula was a vast area with few scattered resources to sustain life: apart from a few choice spots at coasts, oases, or in mountain valleys to sustain agriculture, mankind subsisted in a highly competitive mostly nomadic way of life that made for a very traditional, tribal and feudal society, which has religion and politics fully intertwined.

Then huge oil deposits were struck, and it started a flood of capital and investments, enabling a large growth of inhabitants and infrastructure in an environment that still doesn't support a self-sufficient high-density population. At 1/100 European density in 1950 it has since grown tenfold, but mostly from labor immigrants which have as little political sway as the slaves that were an integral

part of that traditional society, even if outlawed since.

The United Emirates is a federation of absolute monarchies covering only 2.5% of the peninsula along the oil rich north facing coastal strip on the Persian Gulf and its social code has seen the minimum possible evolution from its traditional tribal base, as its emirs see very little value in much of the social code, Western democracies might want to offer or impose.

But very much like China's political elite they understand the importance of controlling the social code AIs incorporate to maintain sovereignty and so they have invested very heavily into large language models, which have been trained with a strong focus on Arabic language inputs, while English is used to augment the general knowledge base [6].

And contrary to China⁶ or OpenAI, UAE's based Group 42 Holding Ltd., largely owned by a son of the UAE founder and brother to the monarch of Abu Dhabi, decided to make them globally available as open source under the permissive Apache 2.0 license. That is especially remarkable, as these 13 and 30 billion parameter JAIS models have been trained on perhaps one of the most modern and powerful AI platforms available today, a Cerebras wafer-scale based supercomputer estimated to cost almost a billion US\$ when completed at the end of 2023 [7]. Fully owned by G42 it runs in Silicon Valley, in a data center just a short walk from Nvidia headquarters.

To recap, the UAE with a population of 10 million, of which 1 million are citizens, and 7 men are emirs holding political power, is so determined to ensure its sovereignty and control over the social code embedded in the large language models used domestically, that it spent \$1000 per citizen on LLM hardware alone.

The EU with a population of 450 million (mostly citizens) does not spend 450 billion Euro on AI, but currently focuses on regulating AI to maintain sovereignty; which is extremely important, but beyond the natural boundaries of your sovereign

⁶ There are also various open source LLM project from China vying for public attention, e.g. 01.ai, which is also supposed to represent \$1B in investments. Currently no documentation is available and first

hands-on tests show questionable quality compared to similar sized JAIS or Llama-2 LLMs

territories, bargaining power requires deeper investments.

The UAE is not renowned for its egalitarian altruism⁷, and given the political context the UAE's JAIS 13B and 30B models may in fact be designed more as weapon of social code hegemony among the global population of 135 million Arab speakers and perhaps even 2 billion Muslims, whose core codex is *read* in Arabic—even by its closest enemies, who used to rule them from across the gulf but prefer to speak Farsi. They all have recognized the dire need for LLMs that actually support their ethos, instead of Western models showing at best ignorance, and often quite a negative bias towards Islam, the Arabic world, or a class society. But Iran, like Russia and North Korea, belongs to a group of nations the West would like to slow even more than China on their path to AI supremacy.

The United States have used their lead and stake in IT for decades of foreign political control, and apart from smuggling and piracy, open source has been one of the few escapes available to contenders. Many started their own Linux distributions, forked digital essentials, and took control over their domestic Internets. And China is only the most capable, when it comes to replacing every bit of IT that the US aims to use as a control lever, many others are just as motivated and search for allies with ever fewer constraints.

With the technology foundations of IT, value system divides seem of little issue. RAM and flash chips, even CPUs and most ASICs remain free of political bias. Operating systems definitely have a somewhat leftist ethos when it comes to scheduling, but it's easily configured more feudal for real-time. Likewise, much of the stack of digital essentials like browsers and document management applications can easily be made to conform to local ethos. Using them as open source across the biggest political trenches should be minimal effort while a global user base helps to spread the development effort [8].

Of course, shared code is a gift that can be weaponized like the very first Trojan horse, the main benefit is that this is more easily discovered than in the closed binary distributions of yore.

⁷ Its moral codex encourages giving alms to beggars, not abolishing their career path or that of emirs e.g. via universal basic income

Sometimes redesigning the very von Neumann base of computing and programming languages may seem attractive to eliminate the fundamental vulnerability that accompanied the invention of software [9]. Especially Linux as the global operating system is straining under the inadequacy of a 1960's design in a world where millions of heterogeneous cores may need to be coordinated to run applications spread across network fabrics that start computing within the data plane and cover the planet. But, as we can observe in nature, evolution practically never leaves enough of a gap for really fundamental reboots, as each iteration needs to survive, first.

That is perhaps the reason why the UAE decided to step in early and with such significant effort to provide an evolutionary branch of LLMs substantial enough to avoid that OpenAI supremacy which would drown out their ethos. And instead of limiting itself to some last-minute and usage constrained leaking like Meta, they seem to have faced that challenge with a very wide, long-term and open-source perspective.

Necessity played a significant part. LLMs need data to train and whereas the openly available text content base for English encompasses around 2 trillion tokens⁸, no other language comes even close. For Arabic as few as 55 billion usable input tokens could be found which limits the optimal model size in accordance to Chinchilla scaling [10] to around 7 billion weights. And those would not cover large scientific bodies with sufficient substance, or many areas of practical human knowledge, because they had never been published in anything but English. Arabic itself first needed to be trained by the model, using an approach very similar to how children learn: exposure or immersion with context.

It was next empowered to understand topoi missing from the Arabic language bodies using English documents translated into that language to enable bilingualism, which could then be used to expand the knowledge base directly from English texts, as if they had been natively written. Most importantly that knowledge can now be rendered in both languages equally, widening the scope of knowledge

⁸ A minimal carrier of meaning, of which there could be several in composite words or with agglutinating grammars

that is directly accessible via the Emirati’s mother tongue⁹.

While this effort has required a significant investment not just into infrastructure but base research, significant parts of that base can be reused for other languages which also have significant grammatical challenges and gaps to overcome between themselves and English. And with the expanded infrastructure available, training times have fallen as low as 21 days for the 13b model¹⁰.

We don’t know if Group 42 will offer LLMs-of-a-given-language-as-a-service, but the EU should quickly consider asking that question.

G42 has proven that this approach is generic enough to be repeatable wherever a significant enough language body exists to train the multilingualism, and its organizational, functional and financial viability. It propels the topic of localized large language models for the sake of sovereignty from abstract deliberations to the rather concrete decision at which level the European Union chooses to collaborate with the UAE and like-minded others, or replicate the approach, hopefully again within the largest possible federation of collaborators and contributors.

J AIS comprises not just the base models, but also chat variants, which have been carefully fine tuned to match the ethos of their authors, who concede that protecting against factually or culturally wrong answers remains a challenge and will require further work. But while the cultural training each derivative of these base models will be diverse, the basic approach, the code and a collection of curated input bodies, and even the infrastructure that performs training and inference can be shared openly, once policies, procedures, controls etc. have been developed, tested and contractually agreed.

Whether it is large language models or others, that also include picture or video capabilities, the ability to respect and incorporate and respect very localized social codes or knowledge bases is essential to increase the value of AI solutions. As Emiratis would be quick to agree, a very loyal artificial slave

sticking strictly to the rules of its owner can be far more valuable than something more akin to an employee with a college degree, but far too many opinions and an outlandish ethos: reliability, trustworthiness and control may be more desirable than a GAI that lacks these qualities.



Jebel Jais mountain, Ras Al Khaima, UAE (Source: jerry 1882 on Adobe Stock)



The J AIS LLM logo

The need to add not just linguistic translations, but to include “grammars of ethos” into the models will grow, and those who we might consider our greatest political contenders, will be most motivated to generalize the issue and thus make it more cost efficient to resolve if we cooperate with them.

And when it comes to translating even between “speech” and “sonar” cultures like that spatial tribe or interacting with sonic-visual species like

⁹ Arab spoken dialects differ vastly, even within the UAE, but a standard dialect called fuṣṣḥāh (فصحى), is used for nearly all publishing and broadcasting

¹⁰ For the 30b model a quote for the training time could not be found. But it seems safe to assume that training started after the 13b publication in August 2023, while the 30b model was ready for download three months later in early November

squid, dolphins and whales, perhaps only AIs will be able to shorten those fundamental gaps in how we look at the very same universe as “flatlanders”, to the point where peaceful coexistence and informational trade is made possible.

Contrary to China’s leaders¹¹, with who the United Emirates share quite a lot of opinions when it comes to freedom of speech, democracy, or equal rights and responsibilities for everyone, G42 has chosen to go with a government driven Open-Source approach, which may catch on as both collaborate more than Cerebras and the US government appreciate [11]. The EU should go much further and seek a very active collaboration and knowledge transfer, that goes as wide and deep into beyond its borders as possible, because the potential benefits grow, as always, with scale.

Knowing where to draw the lines, where to insert safeguards against the inevitable temptation of weaponizing code requires matching research efforts now, negotiations and regulations later, which should be prepared to match the underlying pace of the scientific and technological advances.

Open source is ancient and was since times immemorial a path from war towards trade; which then often led even to marriages, exchange of not just social but also genetic code.

Open source LLMs present a unique opportunity and a clear mandate to act now.

Countering the Giants

Microsoft, Nvidia, Google, and Meta are only the biggest known players, who are betting their fortunes and their future, and those of their investors, on AI becoming a profitable product in the near future.

Microsoft alone has planned \$50 billion for their 2024 data-center build out, which includes than 400.000 Nvidia GPUs already allocated and depreciating. Google is also investing in TPU based build outs into AI infrastructure costing dozens of

billions, while none of these companies break even with public LLM offerings yet.

It has them irrevocably committed as these investments are fully bespoke, almost as much as Bitcoin mining rigs, and cannot be repurposed to run SAP or any other general-purpose workload.

This means these companies are not ready to sit down and calmly debate the ethical use of AI, the risks of GAI, or fundamentally changing how their LLMs might include local preferences and social code variants, because even a month of delay would redshift the profit line, a year of delay create financial havoc.

When Microsoft was faced during Thanksgiving 2023 by the altruistic faction within OpenAI endangering their running freight train by firing Sam Altman, it immediately offered to take on the entire team of lead engineers for any salary and working terms they might wish for, because such derailing, or just a delay of their AI strategy by AI skeptics would have cost vastly more.

None of the players can afford to pause, rethink or renegotiate, because fab volumes have been contracted with huge penalties for plan deviations, millions of chips are either already produced or in the midst of a production run that is edging ever closer to a full year from committing to push the start button to the finished product filling inventory, after processing and assembly steps that involve specialists across the globe who need to plan years ahead to adjust capacity. In the quest to push out the competition by contracting all available capacity, they have made themselves vulnerable to anything they did not foresee. Vulnerability tends to be compensated by defensiveness and escalates to aggression.

The sheer size of their commitment makes them as dangerous as stepping in front of a two-mile freight train as if it were just a cart, with a hand held high because your grocery bag broke on the tracks. It may well be that they are in fact on the wrong track, a track that violates a nations law or even an EU directive, but penalties would need to

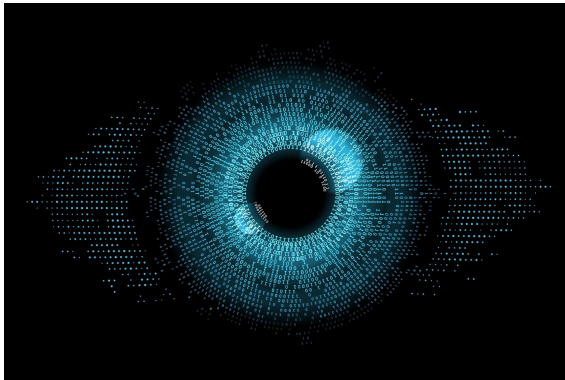
¹¹ While the Chinese government has sponsored local forks of digital essentials such as OpenKylin, the Chinese government is reluctant to push open source at the leading edge. There are quite a few Chinese

LLM projects, increasingly also open source, but so far either company driven or academic.

match the stakes and actually get enforced within the living memory of acting executives.

Stopping the Big Data Steal

Microsoft is pushing AI agents into operating systems and applications that run on Personal Computers: PCs which belong to owners, who as such have a sovereign right for exclusive control. But taking a page from Apple's scrapbook, which developed their products out of a DRM-enabled music player, as well as Google and others, who have invented the data economy, they now have these AIs scrape PCs for any usable data to convert into sellable insights and training material for models they run in their data centers, without asking for permission from the owners of the PCs and the material stored there.



It is the concierge going about your home, reporting back to the recruiting agency on your comings & goings, habits and hobbies, including a carbon copy of any document you might receive, touch or send, in fact even on where you hesitated or rewrote a sentence. As I type this, Word freezes every ten minutes or so, trying to execute a task called ai.exe, describing itself as “Artificial Intelligence (AI) Host for the Microsoft® Windows® Operating System and Platform x64”, a pest I surely never invited into my home and therefore tried to disable, again and again. And no, I’m not “logged in”, nothing cloudy about my machine which operates entirely without a Microsoft account, or that company being permitted anything but software delivery and occasionally unclogging blocked pipes.

Microsoft argues that this nosiness is required “to improve their services and products”. It may also argue that my personal data never leaves my personal computer, or that in case of Office 365, that it actually never left their data centers. They are convinced that the insights that their agent gathers

or their AI then develops are then obviously and exclusively theirs, a court case they urgently need to lose; while neither the watertight regulation, which puts device and data sovereignty solely with the owner, nor that court case have been started yet by the EU.

An AI chat-bot hallucinating may be annoying but seem harmless. However, Microsoft is handing Co-pilot executive power and the steering wheel to our personal computers. Giant investments into making cars sufficiently intelligent for self-driving may have turned their principal imaginator towards hallucinations, but so far have failed to deliver a safe product [12]. Tesla's and Microsoft's need to sell demonstratable AI value is as big as the billions in balance and dwarfs all concerns. Co-pilot's potential for failure and abuse cuts into the core of human value creation in the Internet age and will inspire a completely fresh wave of criminal creativity, where little more than prompting [13] is required to bypass every known security measure invented since the code-data bulwark was breached on June 30th, 1945 [14].

Forty-two years after its first operating system, Microsoft has not managed to make printing documents safe [15] [16]. Currently self-driving Co-pilot only insists on resetting user preferences [17] but forcing it on every desktop [18] without permission or consent...

...clearly needs action.

The sole responsibility of the European Commission is to the citizens of Europe. And their main task is to take actions that benefit them.

What Microsoft, and the other Internet giants racing towards socio-economic dominance via the better, bigger, smarter, and perhaps even super-human AI services, are doing here, does not benefit European citizens in any obvious way that is more significant than the risks involved.

Where they might technically argue that processing all your data via distributed AIs does not represent a violation of the GDPR or other regulation, they are feeding a knowledge base from which any insights you personally developed, can no longer be deleted.

This data extraction must quite simply be stopped, because in a society where content and social code is becoming more valuable than jewels, such thievery needs to be met like the concierge absconding

with your baubles. And it gives these giants an advantage that simply doesn't balance back in EU citizen benefits.

No data or insight from EU personal computing devices may flow into AI giant data centers, unless the owner or his loyal agent hands it over for the purpose of a transaction. No application or operating system may gather data and insights on PCs or their owners by default, nor send it back as "telemetry" or "customer feedback" etc. without the owner explicitly consenting, while no service may be denied if he does not. If payment or license data is required, a neutral clearing house needs to be put in place to safeguard the commercial exchange and contract compliance. Vendors may not demand privileged operational control over a PC to safeguard software purchases, because it's not necessary and done in bad faith.

AI is happening and only time will tell how much good or bad it will do. There is no simple way to undo it from human memory, or stop it from eventually reaching GAI, if that becomes cheap and attractive enough to do. Yet it's hard to imagine it doing so well, without consumers able to buy products or services it helped to create to finance its evolution.

But sovereigns at every level, be they just the owner of a PC or the rulers of nations need, and as sovereigns have every right to control, its spread and usage. And they need to exercise that right in the interest of their citizens or just their families and the ability to do so safeguarded by their government, at least in a democracy.

Such control can't be too complex and it is easier to implement them, if direct financial feedback can be used.

Ensure Local Sovereignty

Sovereign control needs to be decisive, with every default pointing towards the owner of the devices and the data not the hardware or software vendor. A list of topics that should be very openly and noisily discussed to help steer planned AI giant commitments could include:

- vendor locked personal computing devices need to be in fact unlocked by default, the delegation of management to a vendor, a conscious but reversible decision that doesn't imply complete data loss when changed

- Any cloud synchronization or telemetry data option again needs to be off by default, only enabled when the owner agrees with informed, non-discriminative and reversible consent. These cloud synchronization offers may not be vendor exclusive, 3rd party providers enabled via open APIs
- No application that can run with local data storage may prescribe a cloud link to function
- Where such a shared data provider is necessary (e.g. collaborative tools), the storage needs to be geographically local (minimum country level) and possibly 3rd party enabled
- No generalized analysis of cloud data by code or models that is not specifically agreed by the owner: you can continue to pay with your data, but just entering a web site or installing an application may not be interpreted as eternal and all-encompassing consent
- No generalized analysis of local data on personal computing devices or appliances by apps or models, no such modules or packages may come activated by default or be inserted via some update without explicit informed, revocable, and non-discriminative consent
- Applications should never look at data they have not generated or were instructed to use by the owner

Philosophically it means that PCs and appliances may not be turned into agents of an AI or cloud giant: consumers must specifically want to purchase them as such and be clearly informed as to what that means now and whenever that changes. And they must be able to use these devices without the vendor connection, where that makes sense (e.g. coffee machine).

Rather concretely it means Microsoft may not just install Co-Pilot on Windows PCs or convert Office installations into a cloud variant unless the owner gives its informed and reversible consent.

The creeping backdoorism which has been designed to feed the training pipelines needs to be cut before it's proclaimed the universal state of the art. And after that, even more so.

Reigning in the Networks

WhatsApp, Facebook, Twitter etc. have made themselves "indispensable", mostly by just being convenient and using the network effect. It has led

to schools, clubs, employers, service providers, governments etc. using them to the point where they become effectively critical infrastructure or mandatory to participate and function.

It means the EU citizens no longer have a realistic choice not to feed these foreign giants with all the data their apps might siphon from their devices.

That choice not only needs to be reestablished, not feeding them needs to be the preferred default.

And that starts by all branches of EU governments going off social media not operated by themselves or by a 3rd party under their full control. Even if social media were rocket science, after ten years of use and study, far lesser nations have built commodity rockets.

And it then needs to follow through in the private sector, where no employer may rely on social media or AI platforms, that are under foreign sovereign control. E.g. if an employer uses WhatsApp to direct employees, any employee can file an anonymous complaint that is followed up and sanctioned so the employer will have to switch to something like a self-hosted or safe EU regulation audited chat platform, that sticks to chatting.

If citizens want to use TikTok, Facebook or Instagram for their private life, that's their free choice. But when it concerns anything not strictly private, local sovereignty must take precedence.

Facebook could become such a trusted 3rd party operator, if they split operationally along the sovereignty lines with audited functionality and data flows. If they were better and cheaper than a 2nd rate EU implementation, that should benefit EU citizens. Facebook has asked to be regulated, perhaps for that reason, but without such strict rules, corporate profit optimizations can only grow the size of the unregulated freight train [19].

And when it comes to Microsoft, the need to break its stranglehold on the workplace desktop is bigger than ever. All branches of EU government need to switch to open-source alternatives for Microsoft products such as the desktop OS and the office suite and every employee needs to have the right to refuse Office 365, Co-Pilot or anything that uses OpenAI services without discrimination. As Russia, China, North-Korea and others will attest, sovereignty depends on the ability to control the platforms that run code.

And when these use AI models that develop a mind of their own, the control needs to match their rise in capabilities.

Conclusion

As with any tool or technology, AI's benefit to EU citizens depends on its usage.

As with most Information Technology advances, leading edge AI requires scale to distribute the cost of its evolution.

Control vs. scale are not diametrical opposites, when you carefully sort areas of synergies and areas where differentiation and partitioning are required, scale can actually bridge value divides enabling lower cost. But the current players are both driven and tempted too much by short term goals and practical monopolies.

It's much easier to recognize the disruptive potential of how the current AI giants drive LLMs towards GAI than to imagine than how EU citizens will benefit. They seem only concerned about scaling out for affordability and market dominance, while they disregard the need to differentiate and partition.

Yet not only must such platforms support the full diversity of existing social code, they might even need to encourage, but surely enable, a rapid further growth of diversity for conflict avoidance where people come virtually closer and more involved with each other, without wanting to give up their individual or group ethos. Interoperability is key, exerting pressure towards a standardized ethos can have explosive consequences and is a sovereign privilege.

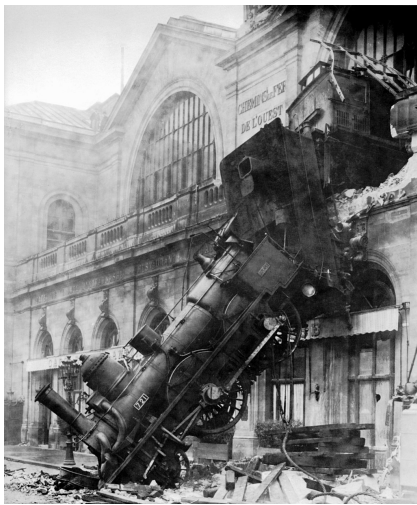
If AI is to become truly ubiquitous, a constantly available and intensively used commodity to empower humanity, one or even a few giant global corporations are the worst place to hold control, which must match the topology of the society on the ground, from federations of nations to the level of an individual.

Where those sovereigns are not ready to take the lead themselves, they must ensure their control in other ways. The right to internalize or nationalize such control is the unalienable right of any sovereign, but corporates may need reminding and courts clear directions to follow citizen over corporate interests.

Scale is the primary means of lowering the cost of technology evolution, but scale exerts gravitational pull when it becomes huge. Before corporate scale overwhelms sovereign power, those need to step in and slice, relentlessly if corporations do not anticipate the need to pre-partition along existing social code diversity.

Without that balance of power AI will fail to benefit those who ultimately enable and support it by paying for services and products which make use of it. And without customers, corporations die.

The EU needs to partner with the widest range of other sovereigns to create a framework of controls, to avoid corporate run-away freight trains accumulating through cost optimizations, which then become too dangerous to operate and steer. Using and enforcing open source and open processes at speeds and power that match the technology underneath is crucial and the EUs immediate obligation with the broadest possible range of partners.



Train wreck at Montparnasse 1895
(Source: Studio Léry and Sons / Wikimedia)

References

- [1] "'Flatland' on Wikipedia," Wikipedia, [Online]. Available: <https://en.wikipedia.org/wiki/Flatland>. [Accessed 20 December 2023].
- [2] T. Hoberg, "Extreme reuse: the only future any code can afford. HiPEAC Vision 2021," 18 January 2021. [Online]. Available: <https://doi.org/10.5281/zenodo.4719690>. [Accessed 20 December 2023].
- [3] "'Descendants of Queen Victoria' on Wikipedia," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Descendants_of_Queen_Victoria. [Accessed 20 December 2023].
- [4] "'Dabbawala' on Wikipedia," Wikipedia, [Online]. Available: <https://en.wikipedia.org/wiki/Dabbawala>. [Accessed 20 December 2023].
- [5] D. Patel, "SemiAnalysis," [Online]. Available: <https://www.semianalysis.com/>. [Accessed 20 December 2023].
- [6] Sengupta, Neha et al, "Jais and Jais-chat: Arabic-Centric Foundation and Instruction-Tuned Open Generative Large Language Models," 29 September 2023. [Online]. Available: <https://arxiv.org/abs/2308.16149>. [Accessed 20 December 2023].
- [7] S. Ward-Foxton, "Cerebras Sells \$100 Million AI Supercomputer, Plans Eight More," EE Times, 20 July 2023. [Online]. Available: <https://www.eetimes.com/cerebras-sells-100-million-ai-supercomputer-plans-8-more/>. [Accessed 20 December 2023].
- [8] T. Hoberg, "Europe's need for digital essentials, individual sovereignty and consumer protection. HiPEAC Vision 2023," 16 January 2023. [Online]. Available: <https://doi.org/10.5281/zenodo.7461962>. [Accessed 20 December 2023].
- [9] T. Hoberg, "Reversing John von Neumann and Steve Jobs, but not software. HiPEAC Vision 2021," 18 January 2021. [Online]. Available: <https://doi.org/10.5281/zenodo.4719396>. [Accessed 20 December 2023].
- [10] Hoffmann, Jordan et al, "Training Compute-Optimal Large Language Models," 29 March 2022. [Online]. Available: <https://arxiv.org/abs/2203.15556>. [Accessed 20 December 2023].
- [11] T. Mann, "After bashing Nvidia for 'arming' China, Cerebras's backer G42 alarms US govt with suspected Beijing ties," The Register, 28 November 2023. [Online]. Available: https://www.theregister.com/2023/11/28/cerebras_g42_china_refile/. [Accessed 20 December 2023].
- [12] E. Helmore, "Judge finds 'reasonable evidence' Tesla knew self-driving tech was defective," The Guardian, 22 November 2023. [Online]. Available: <https://www.theguardian.com/technology/2023/nov/22/tesla-autopilot-defective-lawsuit-musk>. [Accessed 23 December 2023].
- [13] M. Pesce, "Go ahead, let the unknowable security risks of Windows Copilot onto your PC fleet," The Register, 11 October 2023. [Online]. Available: https://www.theregister.com/2023/10/11/microsoft_expects_it_pros_to/. [Accessed 23 December 2023].
- [14] J. von Neumann, "First Draft of a Report on the EDVAC," 30 June 1945. [Online]. Available:

<https://web.mit.edu/STS.035/www/PDFs/edvac.pdf>.
[Accessed 23 December 2023].

- [15] R. Speed, "Messed up metadata could be to blame for Microsoft's Windows printer woes," *The Register*, 8 December 2023. [Online]. Available: https://www.theregister.com/2023/12/08/messed_up_metadata_to_blame/. [Accessed 22 December 2023].
- [16] R. Speed, "Microsoft to kill off third-party printer drivers in Windows," *The Register*, 11 December 2023. [Online]. Available: https://www.theregister.com/2023/09/11/go_native_or_go_home/. [Accessed 8 December 2023].
- [17] R. Speed, "AMD graphics card users report gremlins with Windows 11," *The Register*, 3 October 2023. [Online]. Available: https://www.theregister.com/2023/10/03/amd_graphics_users_report_problems/. [Accessed 23 December 2023].
- [18] "There Will Be A Personal Computer On Every Desk In Every Home - Bill Gates (1990)," YouTube, [Online]. Available: <https://youtu.be/X-M3FbIlqQA>. [Accessed 23 December 2023].
- [19] "'2023 Ohio train derailment' on Wikipedia," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/2023_Ohio_train_derailment. [Accessed 20 December 2023].
- [20] T. Hoberg, "Gaming, content and the metaverse. HiPEAC Vision 2023," 16 January 2023. [Online]. Available: <https://doi.org/10.5281/zenodo.7461953>. [Accessed 20 December 2023].
- [21] "'Standard Oil' on Wikipedia," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Standard_Oil. [Accessed 20 December 2023].

Thomas Hoberg is the technical director
at Worldline Labs, Frankfurt, Germany

Artificial intelligence





Artificial intelligence (AI) and especially generative AI represents a new frontier for innovation and has become a central point of attention since the public release of ChatGPT in November 2022. It is a pivotal point of strategic competition globally. As Europe navigates the challenges and opportunities presented by AI, particularly large language models (LLMs), it is imperative to adopt a coherent and proactive strategy that ensures growth, leadership, and the ethical deployment of these technologies. HiPEAC outlines key recommendations for EU's approach to AI, with a focus on developing AI models aligned with "European" values, which can be executed on edge devices or by a federation of edge devices. The HiPEAC community should develop always more efficient AI accelerators and software but helped by AI-based tools for better productivity in software and hardware developments.

The very fast evolution of artificial intelligence

by various AI-based tools, including ChatGPT from OpenAI¹, Marc Duranton

It is hard to believe that just a few years ago, the reach and accessibility of artificial intelligence were much more limited. Today, AI based on neural networks have made it easier and cheaper to create various content, like texts, pictures and even code. These neural networks come with features like writing, summarizing, text-to-speech, speech-to-text, picture and video creating and editing, transcription and translation, etc. These features are particularly beneficial for creators, enabling them to produce content quickly. AI technologies can convert scripts into speech, enhance text to engage viewers, and even co-create original content. Users can also create avatars from their photos to narrate texts, with the technology ensuring natural lip-syncing and expressive speech. There are hundredth of various tools that emerged in less than a year². Even the domains covered by the HiPEAC community can benefit from the increase of productivity: there are AI tools that helps

developing programs (from the specification to debug) and even some of them can help for developing hardware. These two aspects will be covered in separate articles of this HiPEAC Vision.

The easiness of use of AI-based tools also has drawbacks as it can be used to create fake news, pictures or video quite undisguisable from real ones. They can also be used by hackers to develop new, easy-to-use ways to exploit vulnerabilities (or find how to make bombs, etc). As most AI (using Large Language Models or LLMs) have protections against malevolent use, this leads to hackers trying to find ways to bypass the LLMs' protections, "jailbreaking" them by using more and more sophisticated prompt injection. As the nature of AI based on Neural Networks means that the information is not explicit in an identifiable part of the model, it is very difficult to "erase" sensitive information. Of course, they can be trained with

¹ Slightly helped, prompted and organized by Marc Duranton.

² For example, the web site <https://www.futuretools.io/> lists over 2400 of such tools in various domains

curated datasets without “dangerous” data, but the size of the dataset, and the curation, which is mainly done by humans, make this virtually impossible or too expensive. Fine tuning and other more explicit approaches to sort the input and output are being developed, but they cannot 100% guarantee that the AI are “safe” and that they will not generate output that can be used by malevolent people.

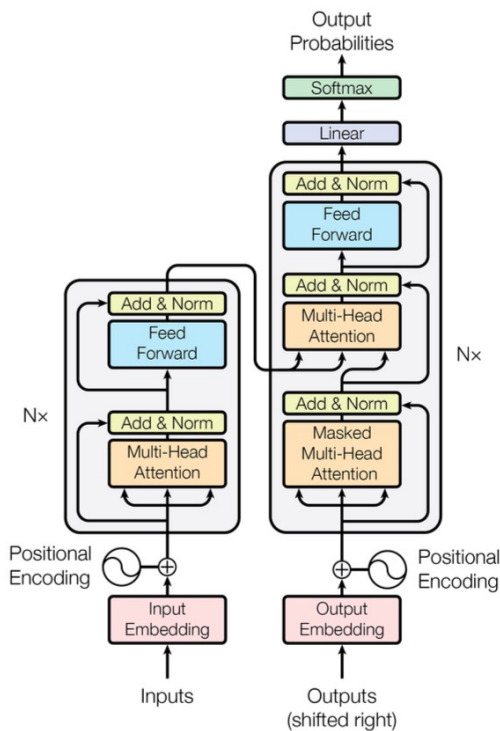


Figure 1: structure of transformers, from [1]

But these recent AI based on the “transformer” model [1] (the goal of these systems is to “predict” the next “tokens” from a series of input “tokens”) can generate possible outputs, but not “real” ones. They are called “hallucinations” when the outputs of such transformer based LLMs (or other large models) don’t reflect the reality. This is why it is very important to check the veracity of the results of those AI. It is easier in some fields (but not always obvious) such as in code generation: the code can be compiled and executed to check if its syntax is correct and if the results are according to the expectations. It can be also inspected manually, or by the help of other AI that can analyse it.

AI technology has become one of the fastest-growing fields, particularly since the introduction of GPT, which first appeared in 2018 but didn’t gain significant momentum until November 2022

with the release of its chatbot version, ChatGPT, offering almost human-like communication and quickly becoming the fastest-growing user application in history. Since then, it has transformed the world, for better or worse. Its success can be explained because it was the first computing system that could do a lot of things without explicit programming or choosing a particular specialized application, making its access easy for anybody who can write (in English mainly) and read.

This era of generative AI also has some impact in the perception of computing systems by people:

- **Enhanced Capability Perception:** Generative AI showcases the ability of computers to not only process data but also to create new content, be it text, images, or even music. This represents a shift from the traditional view of computers as tools for calculation and data processing. This was a shock for some artists and other people in creative jobs who were thinking that they will be shielded from the impact of computing systems.
- **Blurring the Lines Between Human and Machine:** Generative AI challenges our understanding of creativity, traditionally seen as a uniquely human trait. By producing texts, pictures, and other “creative works,” these systems blur the line between human and machine-generated content, leading to new discussions about the nature of creativity and originality. The fact that the AI understands human languages and writes it correctly was also a significant part of the impact of generative AI, as language is also seen as a human trait. The advanced capabilities of generative AI have led to new discussions about what constitutes intelligence, prompting philosophical and scientific discussions about the nature of intelligence itself (and a redesign of IQ tests and other tests supposed to quantify human capabilities).
- **Results from computers (running generative AI programs) are not always correct:** for the first time, people experienced results of (AI-based) programs that are not reflecting the reality, which are the results of “hallucinations” and no more trustable as humans. Before, computers were supposed to always give “good” results (apart from bugs or programming errors, but they “are executing what they are told” by humans).
- **Interactivity and Personalization:** With AI systems capable of generating personalized

content in response to user inputs, the perception of computing systems has evolved from static machines to interactive, responsive entities. This personalization makes technology more accessible and appealing to a broader range of users.

- **Raising Ethical and Societal Questions:** The capabilities of generative AI have sparked conversations about ethics, privacy, authorship, and the potential for misuse. This has led to a perception of computing systems as not just mere tools. As generative AI takes on tasks that were traditionally performed by humans, there is a growing perception of computing systems as potential substitutes or supplements to human labour. This impacts how people view their career paths and the skills they need to develop. People who may not have had the skills or resources to create certain types of content can now do so with the aid of these AI systems, changing the perception of computing from specialist tools to general-purpose enablers of creativity.

The HiPEAC Vision 2023 stated that “today’s large models will be optimized and will be able to run on edge devices in the future due to algorithm improvements, optimization tools (pruning and quantization) and optimized hardware.” At that time, we expected that it would be achieved in several years, but it is already realized at the end of 2023 with models with similar performance³ of ChatGPT 3.5 running on consumer grade computers⁴, and announcements at the end of 2023 from providers of chipsets for smartphones, indicate that those models will run in (high-end) smartphones in 2024. It is clear that thanks to the new algorithms/structures of AI models that get performance comparable to very large models (of 100s of billions of parameters) with models will only 10s of billions of parameters, thanks also to the improvement of AI accelerators, the technology of generative AI can be run on embedded devices, today smartphones, soon smaller systems. We see also ideas of combining more specialized (or fine-tuned) neural network architectures together to get better performance. This is exemplified by the Mixture of Experts⁵ approach, but it might be extended to having several small AI working closely together (in federation, in swarm of AI, similar to the infrastructure proposed in the next Computing

³ For example Solar 10.7B – see <https://huggingface.co/upstage/SOLAR-10.7B-Instruct-v1.0>

⁴ The LLMs models such as Solar 10.7B, LLaMA-2 13B, Mistral 7B or their fine-tuned version can run on a Mac mini with a power

paradigm – NCP – in this HiPEAC Vision 2024). It is important to note that the smaller models which are in competition with the “larger” ones (such as GPT 3.5) are generally fine-tuned versions of “open parameters” models, i.e. foundation models where the parameters are disclosed in accessible and open repositories such as HuggingFace. Compared to January 2023, nobody would have guessed that these open models would have been competitive and so numerous (HuggingFace offers more than 450 000 models⁶ at the time of writing). This was mainly triggered by Meta which released in February 2023 to researchers the foundation models LLaMA, and the pioneering work of Stanford researchers who showed that these foundation models can be fine-tuned for a rather low cost [2].



Figure 2: image created by Dall-E 3

Key insights

- On November 30, 2022, OpenAI released ChatGPT. It is uncommon for a single product to create as much impact on the tech industry as ChatGPT has in just one year.
- Together with other techniques to generate text or images (Midjourney, Dall-E), generative AI made the buzz in 2023, generated a lot of expectations and raised a lot of ethical and societal questions.

consumption of 20W while executing the model.

⁵ Like Mistral.ai’s Mixtral 8x7B - see <https://mistral.ai/news/mixtral-of-experts/>

⁶ See <https://huggingface.co/models> for the actual numbers

- These tools are amplifiers of productivity, and can be used in the HiPEAC fields, such as software development and even for helping creating hardware.
- Open parameter foundations models, often available on HuggingFace, triggered a large development of research and experimentations by various groups, leading to smaller models having competitive performance compared to closed source models such as GPT 3.5.
- These smaller models can be executed already in 2023 on consumer grade devices (PC) with good performance, thanks to open sources developments⁷. LLMs will be in smartphones by 2024, driven by advancements in SoC processors like Qualcomm Snapdragon 8 Gen 3 [3] and MediaTek Dimensity 9300 [4].

Key recommendations

Invest in AI core technologies and development across application domains to support growth for Europe

Europe must invest in research and infrastructure to support multimodal (by integrating text, image, and sound) AI development, ensuring its applicability across sectors such as healthcare, education, and public services, thereby fostering an environment of innovation and practical AI utility.

Develop and provide access to foundational models that support "European" values

Europe should lead the development of foundational AI models. This involves creating and sharing methodologies and datasets for fine-tuning these models to suit specific regional needs. By doing so, Europe can secure its sovereignty in AI technologies and promote a digital economy that reflects its standards and ethics.

Promote open-source models

Europe should encourage the growth of Open-source (open parameters) models which serve as the backbone of a collaborative AI ecosystem,

ensuring open access to AI resources and facilitating a culture of shared progress.

Develop local solutions and specialized accelerators for the integration of large models in smart devices

Europe should lead the developments of making AI based on large models suitable for integration into smart devices at the edge. This empowers real-time AI applications on devices, reducing dependency on centralized data centers, and enhancing privacy and efficiency. Europe's strategy should include support both for start-ups and for established companies in developing edge AI capabilities, fostering a decentralized and resilient AI infrastructure.

Use AI to aid software and hardware development

Europe should improve the productivity of current engineers and researchers by upskilling them with AI, leveraging these technologies to address complex computational problems and accelerate the development cycle. The concept of "centaurs", or partnerships between AI and developers, is a promising approach to enhance productivity and code quality. Europe should invest in training, tools and platforms that facilitate this symbiosis, enabling developers to harness AI for more efficient and creative development processes.

Continue to develop policies around accessibility and societal impact

The societal impact of AI cannot be overstated, and its accessibility and ethics is a cornerstone of digital inclusivity. AI technologies should be available to all European citizens, ensuring that the benefits of digital transformation are equitably distributed.

Ensure "correctness by construction" in AI models

Making sure that AI models generate sound and validated answers (and do not "hallucinate") is key to ensure that AI can be effectively used to carry out the previous recommendation on accessibility and social impact. Europe should lead the way in developing methods that automatically verify the

communities around the LLM inference with more than 390 contributors – it is available at <https://github.com/ggerganov/llama.cpp>

⁷ Especially thanks to the work of Georgi Gerganov who developed the initial LLaMa.cpp. It implements the Meta's LLaMa architecture in efficient C/C++, and it is one of the most dynamic open-source

correctness of AI-generated outputs, thus reducing the need for extensive human oversight and increasing the trustworthiness of AI systems.

The European Union's Artificial Intelligence Act

The European Union's Artificial Intelligence Act (AI Act), initially proposed in 2021 and recently approved at the political level among major EU institutions, will become applicable after a transition period of up to two years. Its primary thrust is to establish a risk-based system for marketable products. AI systems are classified according to the risk they pose to users, with high-risk systems subjected to more stringent regulations and certification. Some techniques, e.g. emotion recognition at the workplace, are banned. As new norms will be formulated by a new regulatory agency at the EU level, the AI Act will begin to have direct implications for software designers, compelling a deeper focus on evaluation, risk assessment, and fundamental rights impact assessment. For generative AI, the Act imposes a set of mandatory measures on all models, focusing mainly on datasets and copyright protection. Providers of highly capable models with systemic risk get extra obligations, including obligatory reporting of technical features and incidents to the Commission. Generated content will have to be identifiable through watermarks, and models will have to undergo independent red-teaming. These obligations pose a significant scientific challenge and mark an important policy change for the research community.

Figure 3: The European Union's Artificial Intelligence Act (from Alexei Grinbaum, CEA)

The rationale of all these recommendations will be detailed in a set of articles. The first one will describe the rapid evolution of generative AI in 2023. The second one will focus on AI assisted software engineering. The third will focus more on the hardware side with the use of AI to help EDA (Electronic Design Automation). The fourth one will deal with the position of Europe in this field and on the on-going development of

regulations about AI, including the European AI Act.

Conclusion

The strategic approach to AI outlined in these recommendations presents a roadmap for Europe to navigate the AI landscape. By embracing these initiatives – for example, developing “made in Europe” foundational models and promoting open science or developing a complete ecosystem of models, hardware accelerators and applications of AI running at the edge - Europe can foster an AI ecosystem that is not only competitive but also reflective of the EU’s commitment to open innovation, ethical standards, and societal wellbeing. By upskilling current engineers and researchers, Europe can amplify their productivity with AI, leveraging these technologies to accelerate the development cycle and improve the position of Europe in the international economic landscape.

To achieve this, and thereby to realize the full potential of AI as a force for good in European society, it will be necessary to coordinate action across all levels of governments, industry, and academia.

References

- [1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, "Attention Is All You Need", [Online]. <https://arxiv.org/abs/1706.03762>. [Accessed 29 November 2023].
- [2] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, Tatsunori B. Hashimoto, "Alpaca: A Strong, Replicable Instruction-Following Model", Stanford University, [Online]. Available: <https://crfm.stanford.edu/2023/03/13/alpaca.html>. [Accessed 29 November 2023].
- [3] "Snapdragon 8 Gen 3 Mobile Platform," [Online]. Available: <https://www.qualcomm.com/products/mobile/snapdragon/smartphones/snapdragon-8-series-mobile-platforms/snapdragon-8-gen-3-mobile-platform>. [Accessed 14 December 2023].
- [4] "MediaTek Dimensity 9300 deep dive: A true Snapdragon rival?," 6 November 2023. [Online]. Available: <https://www.androidauthority.com/mediatek-dimensity-9300-explained-3381678/>. [Accessed 14 December 2023].

Marc Duranton is a researcher in the research and technology department at CEA (the French Atomic Energy Commission) and the coordinator of the HiPEAC Vision 2024.



Artificial Intelligence (AI), particularly in its generative form, has emerged as a groundbreaking area for innovative advancements, drawing significant focus since ChatGPT's public unveiling in November 2022.

The rapid evolution of generative AI in 2023

by various AI-based tools, including Chat-GPT from OpenAI¹, Marc Duranton

The Dawn of Generative AI – Redefining Innovation

Since the last HiPEAC vision, there was an extraordinary boom of artificial intelligence. This was already forecasted in last year edition, but events went even faster than expected.



Figure 1: Explosion of usage of AI in 2023, from LifeArchitect.ai

The availability to the public of chatbots like ChatGPT demonstrated a leap in the quality and coherence of machine-generated text. Such systems could engage in detailed, context-aware conversations, making the interaction far more natural and human-like compared to previous systems. This was really triggered on November 30th, 2022 when ChatGPT was officially released to the public. ChatGPT presented a user-friendly interface

that required no technical expertise to use. Its conversational format made it accessible to a wide range of users, allowing more people to interact with AI technology in a straightforward and intuitive manner. Other companies surfed on this success and proposed alternatives chatbots (such as Anthropic with Claude 2, Google with Bard, ...) and other generative AI application, either to generate pictures (Midjourney, Dall-E, Stable diffusion, ...), or to generate sound or voices, also emerged. This was a tremendous increase in the traffic accessing to all these web sites, as shown in the following figure.

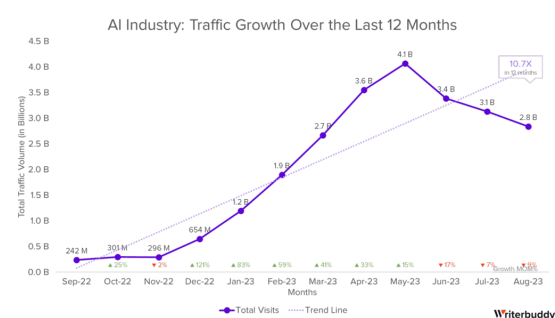


Figure 2: AI Industry: Traffic Growth between September 2022 and August 2023. From <https://writerbuddy.ai/blog/ai-industry-analysis>

The introduction of ChatGPT and similar generative AIs has been a major catalyst for the growth

¹ Slightly helped, prompted and organized by Marc Duranton. Most of the text was generated by Chat-GPT (GPT-4 version). As stated in an earlier HiPEAC Vision, we forecast that at some point in time, the HiPEAC Vision could be mainly written by AI-driven tools. This is an early attempt. We kept the style and repetitions generated by the chatbot on purpose.

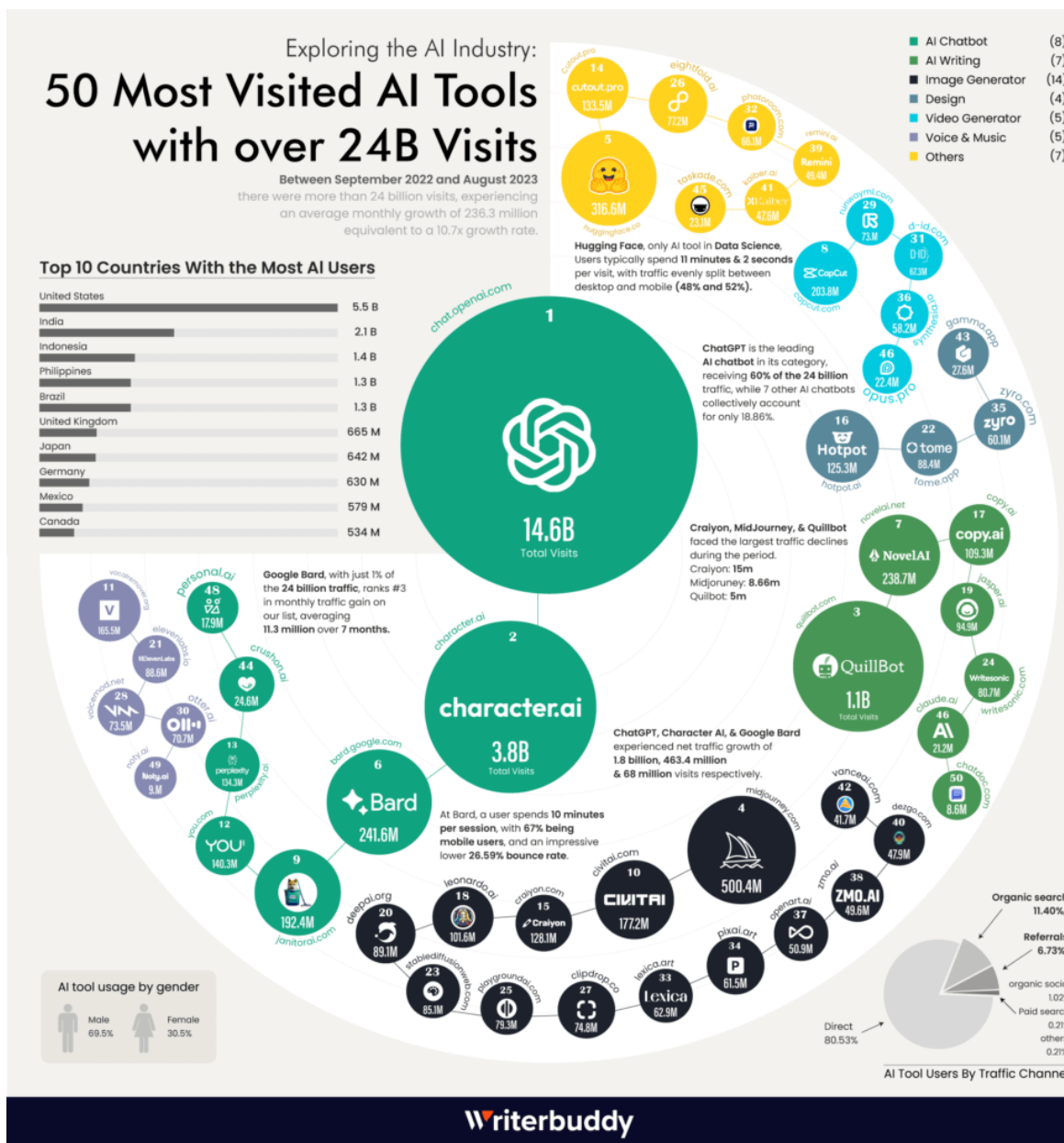


Figure 3: The 50 Most Visited AI Tools and Their 24B+ Traffic Behavior. From <https://writerbuddy.ai/blog/ai-industry-analysis>

and innovation of numerous startups and established companies. Their impact is most evident in customer service, where they powers chatbots that enhance customer experiences through 24/7 support and personalized interactions, leading to improved efficiency and cost savings. In content creation and digital marketing, ChatGPT is utilized by services for generating creative content and marketing copy, streamlining processes for businesses. In the educational sector, ChatGPT aids in creating interactive learning platforms and tutoring services, offering personalized and engaging learning experiences. Furthermore, in the financial sector, ChatGPT and similar systems are

integrated into fintech applications for automated financial advising and fraud detection.

It's widely recognized that the usage of AI has been rapidly increasing across various sectors and applications. This surge is driven by advancements in AI technologies, greater accessibility of AI tools, and the growing integration of AI into everyday devices and services.

The introduction by OpenAI of GPTs (custom versions of ChatGPT made by users) and of the GPT store could jeopardize the business of some companies and start-ups.

In the HiPEAC Vision 2023, it was forecasted that large language models will be running at the edge, and now it is already possible and we can even expect that in 2024 they will be in smartphones. Let's see in a short history how all this happened.

Short history of the recent progress of artificial intelligence

In the realm of technology and innovation, Artificial Intelligence (AI) has long been a subject of fascination and relentless pursuit. It had several ups and downs (“winters”) since the creation of the term “Artificial Intelligence” which was officially introduced by John McCarthy in 1956 during the Dartmouth Conference. In 2012, a significant event in the field of Artificial Intelligence (AI) and machine learning occurred under the supervision of Geoffrey Hinton, a pioneer in the field of neural networks. That year, Hinton and his team made a groundbreaking advancement in the development of deep learning techniques, particularly through their work on a deep neural network. The critical event was the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Hinton's team, consisting of Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton himself, participated in this challenge with a model called “AlexNet”[1]. This model was a deep convolutional neural network that dramatically outperformed other competitors in the competition (AlexNet achieved a top-5 error rate of approximately 15.3%, which was significantly lower than the 26.2% error rate of the second-best entry in the competition). The success of AlexNet in the ImageNet challenge was a pivotal moment for the field of deep learning. It showcased the effectiveness of deep neural networks in practical applications, particularly in tasks involving visual recognition, and marked a turning point in AI research, leading to a surge in deep learning applications across various domains, including speech recognition, natural language processing, and more.

There are several reasons why this was the right moment for the rebirth of artificial intelligence: 1) the algorithm and topology of the Convolutional Neural Networks, 2) the availability of large database for learning. This was done using “supervised learning” meaning that for each input image, there is a label that explains what is this image. 3) The computing power was available. The training was done on a GPU (NVIDIA GeForce GTX 580) and Alexnet required 262 PetaFLOPS for its 61M parameters.

Another seminal moment in AI research was the publication in 2017 of a landmark paper titled “Attention Is All You Need”[2], introducing the Transformer model, a novel approach to neural network architecture that significantly advanced the field of natural language processing (NLP) and machine learning. Authored by researchers at Google, including Ashish Vaswani and others, this paper presented a new method that departed from the then-standard approaches based on recurrent neural networks (RNNs) and convolutional neural networks (CNNs) for processing sequential data, such as text. The key contributions and impacts of this paper include:

- **Introduction of the Transformer model:** The paper introduced the Transformer, a type of model that relies entirely on attention mechanisms, eschewing the use of recurrent layers. This was a significant departure from existing models that processed data sequentially. The Transformer instead allowed for much more parallel processing, leading to faster and more efficient training of large models.
- **Self-Attention mechanism:** A core component of the Transformer is the self-attention mechanism. This allows the model to weigh the importance of different parts of the input data differently, making it highly effective for understanding the context and relationships within the data, especially in language tasks.
- **Impact on NLP (Natural Language Processing) and beyond:** The Transformer model rapidly became a foundational architecture in NLP, leading to the development of highly successful models like BERT, GPT, and others. These models have set new standards in a wide range of NLP tasks, such as language translation, question-answering, and text generation.
- **Enabling large-scale models:** The efficiency and scalability of the Transformer architecture made it possible to train much larger models than before. This has been a key factor in the recent trend towards training very large language models that can understand and generate human language with unprecedented fluency and accuracy.
- **Influence across AI fields:** Although initially designed for NLP tasks, the Transformer model's architecture has influenced other areas of AI as well, including image processing and speech recognition. Its ability to handle

Model	Architecture	Parameter count	Training data	Release date	Training cost	Compute requirement
GPT-1	12-level, 12-headed Transformer decoder (no encoder), followed by linear-softmax.	117 million	BookCorpus: 4.5 GB of text, from 7000 unpublished books of various genres.	June 11, 2018	"1 month on 8 GPUs", or 1.7e19 FLOP.	
GPT-2	GPT-1, but with modified normalization	1.5 billion	WebText: 40 GB of text, 8 million documents, from 45 million webpages upvoted on Reddit.	February 14, 2019 (initial/limited version) and November 5, 2019 (full version)	"tens of petaflop/s-day", or 1.5e21 FLOP.	~ x 88
GPT-3	GPT-2, but with modification to allow larger scaling	175 billion	499 Billion tokens consisting of CommonCrawl (570 GB), WebText, English Wikipedia, and two books corpora (Books1 and Books2).	May 28, 2020	3640 petaflop/s-day, or 3.2e23 FLOP.	~ x 213
GPT-3.5	Undisclosed	175 billion	Undisclosed	March 15, 2022	Undisclosed	
ChatGPT	Undisclosed	? (rumor 20M???)		November 20, 2022		~ x 65
GPT-4	Also trained with both text prediction and RLHF; accepts both text and images as input. Further details are not public.	Undisclosed (1.8 trillion aka 1.8e12)	Undisclosed (13 trillion tokens, aka 1.3e13)	March 14, 2023	Undisclosed. Estimated 2.1e25 FLOP.	

Figure 4: Evolution of Generative Pre-trained Transformers (GPT) in OpenAI (from https://en.wikipedia.org/wiki/Generative_pre-trained_transformer)

sequential data effectively has made it a versatile tool in the AI toolkit.

One year after the publication of Google's paper, the US company OpenAI² created its first Generative Pre-trained Transformer (GPT in short), followed by GPT-2 in 2019, GPT-3 in 2020 etc.

Let's provide a short history of OpenAI's developments. The first model, GPT-1, was developed a year after the seminal paper on the Transformer architecture was published and released on June 11, 2018. GPT-1 was trained over one month using eight GPUs and had 117 million parameters.

Following GPT-1, GPT-2 was released on February 14, 2019, with a significant increase in computing power, about 88 times greater than its predecessor. GPT-2 comprised 1.5 billion parameters.

Then came GPT-3, released on May 28, 2020. It required 213 times more computational power than GPT-2 and had 174 billion parameters. The training of GPT-3 involved nearly 500 billion tokens, using extensive datasets including Common Crawl (with 570 GB of text), English Wikipedia, and several corpora of books.

An evolution of GPT-3, known as GPT-3.5, was released on March 15, 2022. Subsequently,

ChatGPT, a fine-tuned version of GPT-3.5, was introduced on November 20, 2022. This release marked the beginning of a new era for large language models, making this advanced technology more accessible to the public.

Finally, on March 14, 2023, OpenAI released GPT-4. This model is estimated to take approximately 65 times more processing power than GPT-3. Its architecture, though undisclosed, is speculated to contain about 1.8 trillion parameters. It represents a mixture of expert architectures, implying the collaboration of 16 different networks.

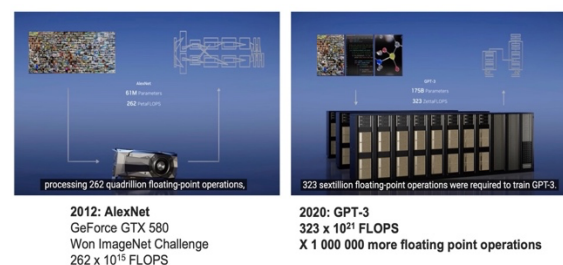


Figure 5: Computing power is driving the advance of ai; from GTC 2023 keynote with NVIDIA CEA Jensen Huang.

² OpenAI was founded in 2015 by Ilya Sutskever, Greg Brockman, Trevor Blackwell, Vicki Cheung, Andrej Karpathy, Durk Kingma, Jessica Livingston, John Schulman, Pamela Vagata, and Wojciech

Zaremba, with Sam Altman and Elon Musk serving as the initial board members.

The landscape of AI underwent a transformative shift with the advent of generative AI represented by Chat-GPT for chatbot, and tools like Dall-E or Stable Diffusion for creating images from text prompt. This new frontier in AI became a particular focal point of global attention following the public release of ChatGPT, which was a fine-tuned version of OpenAI's large language model and served as a general-purpose chatbot. This launch represented a significant step in AI technology, as ChatGPT was able to engage in a wide range of topics and showcase the potential of generative AI in producing human-like text responses. The introduction of ChatGPT marked a shift in public perception and use of AI chatbots. Before its launch, AI chatbots were not widely regarded as highly functional or reliable (even if Chat-GPT is also generating plausible answers, but not real – it is said it “hallucinates”). However, the capabilities demonstrated by ChatGPT in understanding and generating human-like text quickly garnered widespread attention and user adoption. By January 2023, ChatGPT had surpassed 100 million monthly active users, becoming the fastest-growing app in history at that time, outpacing even major platforms like TikTok and Instagram. This period also saw significant developments in the use of generative AI, as ChatGPT showcased its potential in various applications, from helping with homework to assisting in job applications and even drafting political speeches. Its versatility and user-friendly interface contributed to its rapid adoption and integration into daily life for many users. The success of ChatGPT also influenced major tech companies. Microsoft, for instance, invested significantly in OpenAI and integrated GPT technology into various products, including Bing and Teams.

Generative AI, a subset of artificial intelligence, is distinguished by its ability to create novel content, ranging from text and images to complex code, by learning from vast datasets. This capability also democratized the accessibility of AI-powered tools. The transformative leap can be largely attributed to the advancements in neural network technologies. Neural networks, inspired by the human brain's structure and functioning, have become the bedrock of modern AI systems. These sophisticated models, trained on extensive datasets, can analyze patterns, make predictions, and generate content with remarkable accuracy and efficiency³.

³ The reader should have noted that this part was generated by Chat-GPT itself.

OpenAI in 2023

The chronology of OpenAI in 2023 includes global debates on GPT usage in businesses and institutions to avoid plagiarism. In 2023, OpenAI embarked on a tour to promote GPT, addressing criticisms and concerns about the technology. Meanwhile, Italy became the first country to temporarily ban GPT over privacy concerns. That same year, OpenAI released Chat GPT plug-ins, continuing to attract a global user base. The AI industry heated up with competition from companies like Google DeepMind (with Gemini), Anthropic (with Claude 2), who are developing their own versions of GPT. OpenAI faced copyright infringement lawsuits and had to limit free access to paid content. Despite these challenges, the company's annual revenue approached \$1 billion.

In September 2023, OpenAI removed access limitations, but it became evident that distinguishing between GPT-created and human-created content was challenging. That same month, authors like George Martin sued AI for copyright infringement. Alibaba also opened up its AI model to the public.

By October 2023, Chat GPT Plus and subscriptions offered enhanced image generation capabilities, displaying more accurate and detailed visuals. On the anniversary of GPT, OpenAI showcased its advancements, introducing GPTs, where users can now create custom versions of ChatGPT that combine instructions, extra knowledge, and any combination of skills.

On November 17, 2023, OpenAI's board of directors fired CEO and co-founder Sam Altman, but faced internal backlash. Shortly after, nearly all of OpenAI's 770 employees demanded the board's resignation and threatened to leave unless Altman was reinstated. Five days after Altman's firing, the board reached an agreement to rehire Altman, along with the replacement of most board members. One potential reason of

these events is that OpenAI was initially founded as a non-profit in 2015 for AI research, particularly focusing on artificial general intelligence (AGI), OpenAI later evolved. In 2019, it formed a subsidiary to attract more investment and accelerate research, thus becoming a hybrid organization with both non-profit and for-profit dimensions.

Recently, Chat-GPT began to show signs of laziness and stupidity, but developers denied any issues, attributing it to the AI's unpredictable behavior, to user behavior or increased visibility of flaws. The AI seemed to procrastinate, avoid tasks, and even suggest users complete the tasks themselves. Open parameters models, such as Mixtral 8x7B or Solar 10.7B often reach the same level or performance than GPT3.5 on some benchmarks, while GPT-4 performances are still unsurpassed.

Looking forward, OpenAI has trademarked GPT-6 and GPT-7, indicating ongoing development and innovation in the

field. However, predicting the capabilities and impacts of these new iterations remains speculative. The anticipation for GPT-5 is high, with expectations of surpassing human-like communication and analytical abilities.

What are foundation models, fine-tuning, and how Large Language Models work?

A language model is a set of rules that enable understanding and generating text in a specific language, such as English. It operates by calculating the likelihood of each word following the previous one in a sentence.

The language model itself does not produce text independently. Instead, it generates probabilities for each word in the vocabulary. To use a language model, the text must first be broken down into semantic units called “tokens,” which can be words or parts of words. These tokens are then converted into numerical vectors, allowing the model to process them.

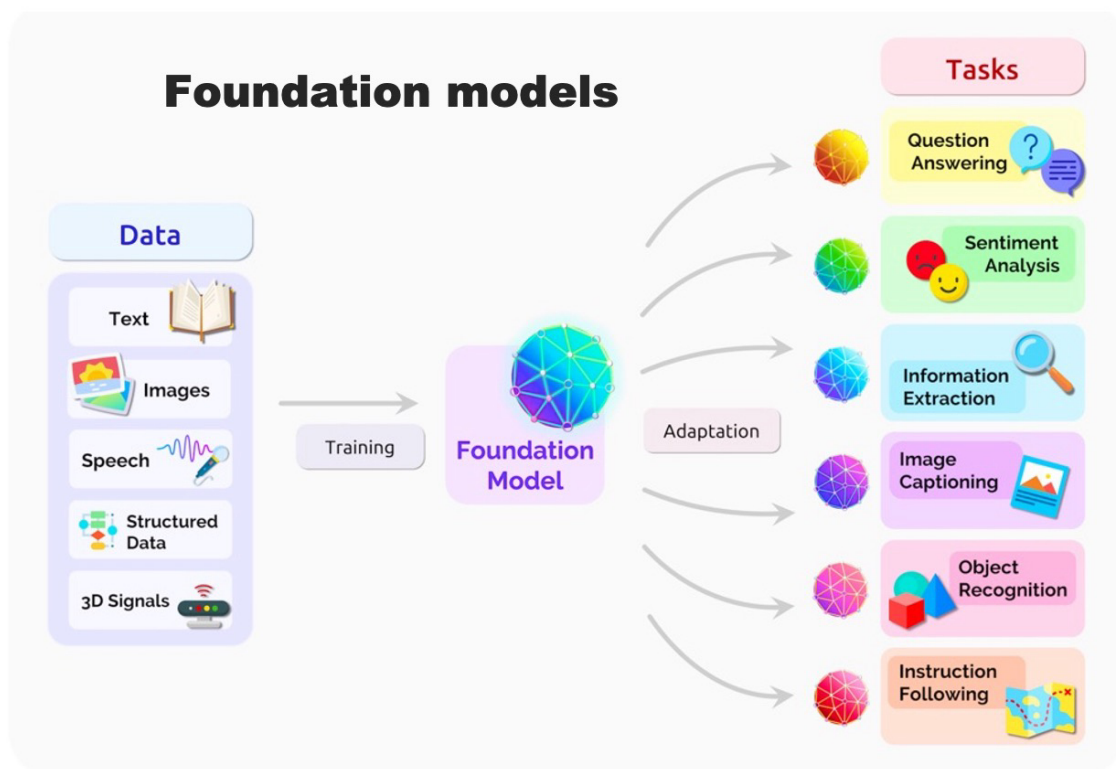


Figure 6 : « A foundation model can centralize the information from all the data from various modalities. This one model can then be adapted to a wide range of downstream tasks. » From « On the Opportunities and Risks of Foundation Models » <https://arxiv.org/abs/2108.07258>

The “context window” size of the model, which is the amount of text memory it can use for understanding, varies depending on the specific model. Then, a process known as “vectorization” converts the text into numerical vectors for the model to process.

Once the text is transformed into vectors, the model generates text using a “sampler.” The sampler selects words based on their probability, but the “temperature” can be adjusted to make the generation more creative or more predictable. This temperature setting controls the randomness in the word selection process, balancing between coherent but predictable text and more diverse but less predictable outputs.

The Transformer model enables the creation of large language models, which are more advanced than traditional foundation models. Let’s quickly define what a foundation model is: these are models trained on extensive databases of texts, images, and various other data types. This training method is known as a self-supervised approach, which differs from the one used in conventional convolutional neural networks. In this approach, the model is expected to predict what comes next in a sequence of items. This is achieved through self-supervised learning, where parts of the data are hidden, and the model must guess, for instance,

the end of a sentence. It can self-correct by revealing the hidden part of the sentence, allowing it to fine-tune its parameters locally, hence the term ‘self-supervised.’

This breakthrough in model design, especially with Transformer-based models, eliminates the need to explicitly label the entire database (unlike for Convolutional Neural Networks – CNN-). Foundation models are created by feeding a large database into Transformer-based models. These models are not initially designed for specific applications; they accumulate knowledge but are not fine-tuned for specific tasks. They can only centralize, “compress” information from various data sources and modalities. After the resource-intensive training phase, these foundation models are adapted for specific tasks, such as functioning as a chatbot, extracting information, answering questions, or recognizing objects. This adaptation process is known as fine-tuning. Fine tuning is far less intensive than the initial creation of the foundation model, and only requires a smaller set of data.

Taking ChatGPT as an example, it was a fine-tuned version of GPT-3 using a method called Reinforcement Learning with Human Feedback (RLHF)⁴. In this process, a prompt is taken from a dataset, and a human labeler demonstrates the desired output behavior. This data then informs

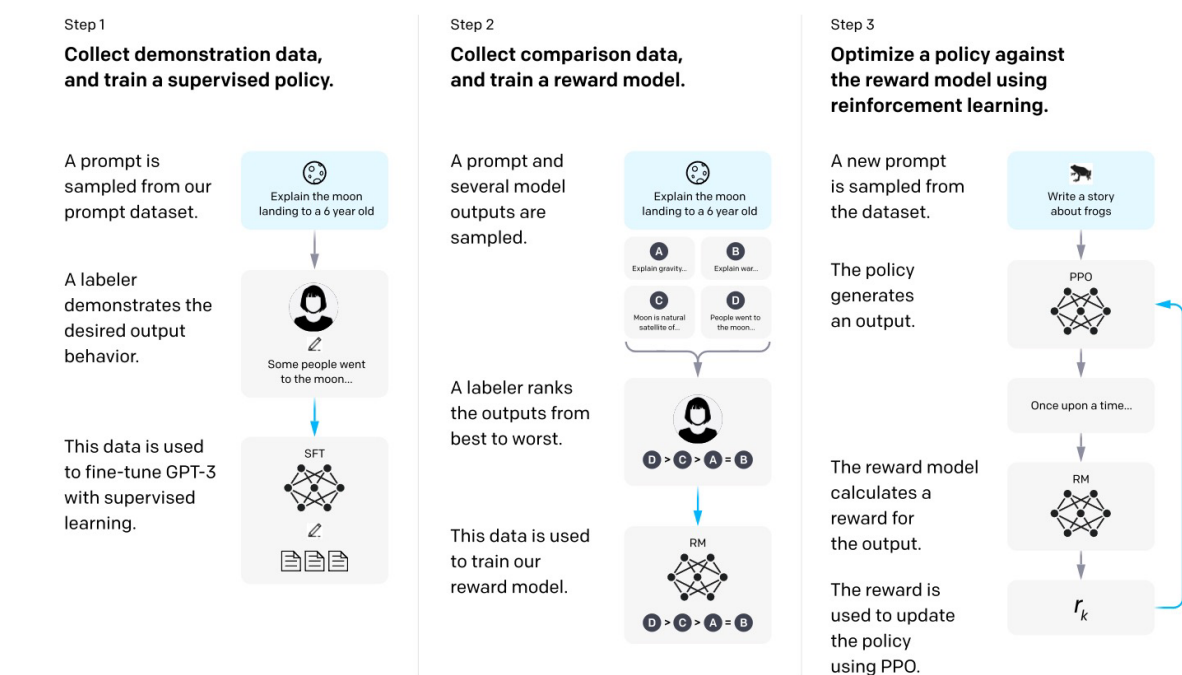


Figure 7: Reinforcement Learning with Human Feedback (used for ChatGPT), from <https://openai.com/research/instruction-following>

the neural network, in this case, GPT-3, through supervised learning. The network generates several outputs, which are then ranked by human labelers from best to worst, and this data is used to further train the model. The final step involves a loop where the network generates outputs evaluated by a reward model. This model calculates rewards for the outputs, and these rewards are used to update the policy. The human labeling in the first and second steps requires significant human resources, and OpenAI outsourced some of this work to a San Francisco-based firm called Sama, which employed workers in Kenya, Uganda, and India to label data for various Silicon Valley clients including Google, Meta, and Microsoft.

LoRA, which stands for Low Rank Adaptation[3], is an advanced method for fine-tuning Large Language Models (LLMs) more efficiently. Unlike traditional fine-tuning approaches that require adjusting all layers of a neural network, LoRA focuses on fine-tuning smaller matrices that approximate the larger weight matrix of the pre-trained LLM. This method has proven to be effective, sometimes even outperforming full fine-tuning, as it avoids issues like catastrophic forgetting, where the pre-trained model loses its original knowledge during the fine-tuning process.

One of the key benefits of LoRA is that it's computationally less demanding. Full fine-tuning of LLMs can be resource-intensive and time-consuming, making it a bottleneck for adapting these models to specific tasks. LoRA's approach is more efficient, leading to faster training and lower computational requirements. For instance, it's possible to create a fully fine-tuned model on a GPU with only limited amount of VRAM using LoRA. This was demonstrated by a team in Stanford that fine-tuned Meta's LLaMA foundation model "*We introduce Alpaca 7B, a model fine-tuned from the LLaMA 7B model on 52K instruction-following demonstrations. On our preliminary evaluation of single-turn instruction following, Alpaca behaves qualitatively similarly to OpenAI's text-davinci-003, while being surprisingly small and easy/cheap to reproduce (<600\$)*"[4].

Furthermore, the trained weights from LoRA fine-tuning are significantly smaller. Since the original model remains frozen and only new layers are trained, the weights for these layers can be saved as a single, much smaller file. This makes it easier and more cost-effective to share fine-tuned models.

In practice, LoRA has been implemented in Hugging Face's Parameter Efficient Fine-Tuning (PEFT) library, which simplifies the process of using this technique. For instance, in the context of supervised fine-tuning, a pre-trained model can be further trained to generate text based on provided prompts, following a format that pairs prompts with responses. In essence, LoRA represents a significant step forward in making the fine-tuning of large language models more accessible and efficient, opening up new possibilities for customization and application in various fields.

Furthermore, foundation models demonstrated what's known as few-shot or zero-shot learning, meaning they can perform tasks with little or no specific training on those tasks, without the need for extensive fine-tuning or adaptation of their parameters, only by adapted prompting.

Zero-Shot Learning: In zero-shot learning, the LLM is able to understand and perform a task it has never explicitly been trained to do. This is possible because of the extensive general knowledge the model has acquired during its initial training on a diverse and broad dataset. When presented with a new task, the model uses its understanding of language and context to infer what is being asked and how to respond appropriately. For example, a zero-shot learning model could be asked to summarize a text or answer a question about a topic it was never specifically trained on, and it would use its general understanding to attempt realizing the task.

Few-Shot Learning: Few-shot learning refers to the model's ability to learn a new task from a very small number of examples. Unlike traditional machine learning models that require large datasets to learn a new task, LLMs can understand and perform new tasks with just a few examples. This is often done by presenting the model with a prompt that includes a few examples of the task being performed. For instance, if you wanted a model to write poems in a certain style, you might show it two or three examples of such poems, and it would generate similar content.

These properties are key elements in the success of LLMs, both zero-shot and few-shot learning demonstrate the versatility and adaptability of LLMs. These models can apply their extensive pre-existing knowledge to new situations, making them powerful tools for a wide range of applications where gathering large, task-specific datasets is impractical or impossible.

Computing power is the fuel of generative AI

As of July 2023⁵, GPT-4 has been reported to have 1.8 trillion parameters distributed across 120 layers, making it more than ten times larger than its predecessor, GPT-3. OpenAI seems to have utilized 16 separate models for GPT-4's transformer module, each with approximately 111 billion parameters. The training of GPT-4 involved around 13 trillion tokens, encompassing both text-based and code-based data, and included some degree of fine-tuning. The cost for training GPT-4 was estimated at around \$63 million, considering the computational power required and the duration of the training period. For inference purposes, GPT-4 operates on a cluster of 128 GPUs, employing techniques such as eight-way tensor parallelism and 16-way pipeline parallelism. GPT-4 also includes a vision encoder, which adds the capability to read webpages and transcribe images and videos. This is in addition to the text processing unit.

From these numbers, it is clear that the evolution of computing power is crucial in the advancement of AI, from convolutional neural networks to large language models. The transition from AlexNet to GPT-3, for instance, highlights the growing demand for computational resources. More optimized computing architectures lead to greater efficiency; for example, moving from general-purpose CPUs to specialized GPUs has resulted in significant gains in operations per energy unit, and further specialization could yield even more efficiency.

Google's development of specialized AI architecture, known as the Tensor Processing Unit (TPU), is a case of more efficiency by more specialization. The first TPU (TPUv1) was announced in 2016, built on a 28 nm process with a processing power of 92 tera operations per second on integer-8 operations and a power consumption of about 40 watts. The TPUv2, announced in May 2017, utilized 20 nm technology, achieving 45 tera operations per second and supporting a special numerical format specialized for AI known as bfloat16, with an estimated power consumption of 200 to 250 watts. The TPUv3, released in May 2018, also supported bfloat16 and had an estimated power consumption of about 200 watts for 90 TOPS⁶. Google continued to innovate with TPUv4, offering 2.5 to 3.5 times more performance than

TPUv3, and the cloud-based TPUv5, which is more efficient and scalable than its predecessor.

NVIDIA's GPUs have become a central architecture in the world of artificial neural networks. The company began as early as 2016 to expand the functionality of their GPUs beyond just graphics and gaming, utilizing their CUDA programming system. This was triggered in 2012, when a significant advancement occurred when researchers successfully employed (NVIDIA) GPUs to reach human-level accuracy in visual recognition tasks (this was ImageNet – Large Scale Visual Recognition Challenge (ILSVRC) as explained in the beginning of this article). This achievement was a harbinger of more recent innovations, like creating images based on textual descriptions. In response to this evolving field, NVIDIA's CEO, Mr. Huang, emphasized in a commencement address at National Taiwan University⁷ that the company had redirected at that time every facet of its operations to support and progress this emerging domain. This innovation demonstrated that GPUs could be effectively used for neural network development. Since then, NVIDIA has continuously enhanced its GPUs to support neural network-based artificial intelligence, including both convolutional neural networks and transformer-based systems.

One of their significant releases was the H100 GPU, which offered peak performance of 2000 teraflops for floating point 16 operations and up to 4000 TFLOPS for integer-8 running on their Tensor Cores. The NVIDIA Hopper H100 Tensor Core GPU is fabricated using TSMC's Custom NVIDIA 4N FinFET Process. The H100 GPU marked a substantial increase in performance capabilities. For instance, training a Mixture of Expert transformer with 395 billion parameters on a dataset of 1T tokens took only 20 hours on (8 000) H100, compared to 7 days with the previous generation A100.

In 2023, Nvidia announced the H200 GPU based on the Hopper architecture. The H200 is a powerful GPU offering 141 GB of HBM3e memory at 4.8 TB/s, nearly doubling the capacity of the previous generation H100. This increase in memory bandwidth allows for significantly improved performance. For example, the H100 GPU was observed to be 11 times faster than its predecessors A100 in tasks like inference with GPT-3 models

⁵ According to <https://www.semianalysis.com/p/gpt-4-architecture-infrastructure>

⁶ According to <https://www.nextplatform.com/2018/05/10/tearing-apart-googles-tpu-3-0-ai-coprocessor/>

⁷ https://www.youtube.com/watch?v=__Ewkal7s3g

Nominal node		28nm	22nm	20nm	10nm	7nm (HD/HP)	5nm (HD/HP)	3nm
Intel	Lg (nm)		24		16	~14nm		
	Fin Pitch (nm)		60 FinFET		FinFET 34 (Now called 7nm)	FinFET (Now called 4nm and 3nm for HP)		
	CPP (nm)		90		54	47		
	M1 (nm)		80		44	26		
	SRAM		HD 0.092µm ²		0.0312µm ²	0.027µm ²		
	Year Publication		VLSI 2012		IEDM 2017/ISSCC2018			
	Risk Prod		2011		1Q18	1H2021		
Samsung	Lg (nm)	32		25	~20	~16	~16	~13
	Fin Pitch (nm)	BULK		BULK	Single Fin 42	FinFET 7LPP	FinFET 5LPE	Horizontal Nanosheets (HNS)
	CPP (nm)	114		86	68	54/57	54/57	40
	M1 (nm)	90		64	51	36	36	32
	SRAM	0.152µm ²		0.084µm ²	0.04µm ²	HD 6T SRAM 0.026µm ²		
	Year Publication	ICSIST 2011		VLSI 2012	ISSCC/VLSI 2017	VLSI 2017/ISSCC2017-2018		
	Risk Prod	2011		2013	1Q2017	1H-19	2H2020	
TSMC	Lg (nm)	30	30	30	25	18	16	~13
	Fin Pitch (nm)	BULK	BULK	BULK	FinFET	FF (30/30)	FF N5 (28/28)	Horizontal Nanowire (HNW)
	CPP (nm)	118	105	90	64	57/64	51/57	45
	M1 (nm)	90	80	64	42	40/40	35/40 (M0 28)	22
	SRAM	0.155µm ²	0.155µm ²		0.03µm ²	0.027µm ²	0.021µm ² / 0.025	
	Year Publication	VLSI 2012	VLSI 2012	VLSI 2014	VLSI 2016	IEDM 2016	ISCC 2020	
	Risk Prod	2011	2018	2013	4Q2016	3Q-17	1H2020	

Figure 8: Relation between the name of the technology node and its real size (from C.Reita, C.Fenouillet-Beranger – CEA-LETI – 2023). What NVIDIA called 4N process is supposed to be a variation of TSMC 5 nm.

with 174 billion parameters. Looking ahead to 2024, it's anticipated that the upcoming H200 GPU will offer 18 times the speed of the A100, reflecting almost a 20-fold increase in performance over three years⁸.

The market of accelerators chips for IA (mainly at the server side) is very important, and several companies are in competition with NVIDIA (and also current customers of NVIDIA's GPU are developing their own accelerator chips, like Amazon's Inferentia and Trainium2, Microsoft's Maia, Google's TPU, Alibaba's Hanguang 800, ...), such as Cerebras (making wafer scale accelerators), Grog, _etched, Graphcore, Gyrfalcon, IBM's Northpole, AMD's Mi300, Intel's Gaudi3, etc...

The market estimate for AI accelerator chips from 2024 to 2030 indicates significant growth. According to Deloitte Insights⁹, the market for specialized chips optimized for generative AI is projected to exceed US\$50 billion in 2024. This forecast represents a major increase from near-zero levels in 2022 and is expected to account for a substantial portion of all AI chip sales in that year. Further, the total AI chip sales in 2024 are predicted to constitute 11% of the global chip market, estimated at US\$576 billion.

Looking ahead to 2027, the AI chip market forecasts vary, with predictions ranging from US\$110 billion to a more aggressive US\$400 billion. However, there are viewpoints that suggest the more conservative estimates may be closer to reality,

given various factors influencing the market. These estimates highlight the growing importance and demand for AI accelerator chips, driven by advancements in AI applications and technologies. However, it's important to note that these projections are subject to change based on market dynamics, technological developments, and other influencing factors.

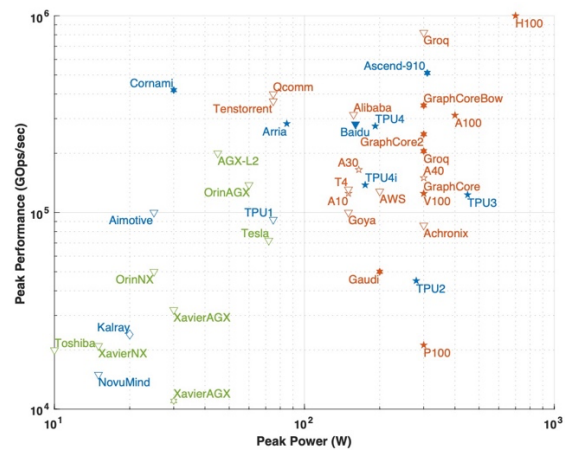


Figure 9: Peak performance vs. power scatter plot of publicly announced AI accelerators and processors (Oct. 2022). From [4].

⁸ From <https://www.nvidia.com/en-us/data-center/h200/>

⁹ ChatGPT gave the following reference: <https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2024/generative-ai-chip-market-to-reach-40-billion-in-2024.html>

The emergence of open source (parameters) foundation models

Up until now, Large Language Models (LLMs) have primarily been running on cloud servers using Nvidia GPUs or similar architectures. This includes more specialized architectures from various competitors. This is mainly due to the large size of foundation models. This is indeed necessary for creating the foundation models as their learning phase requires a tremendous amount of computing and a fast access to a very large set of information. It is still the trend to have a model with more parameters to increase its capabilities (but we will see that smaller models, with different architectures – like mixture of experts – could also have comparable performances on tests and benchmarks than models having 10x more parameters).

Cloud is also the main candidate for inference (use) of the fine-tuned versions of the foundation models, such as ChatGPT. However, hosting these large language models in the cloud has certain drawbacks. One major issue is data confidentiality: transferring data to the cloud raises privacy concerns. For instance, there were incidents where engineers from companies like Samsung allegedly shared sensitive corporate data with chatbot services like ChatGPT. Additionally, cloud server availability isn't always guaranteed, limiting the

accessibility and request capacity for advanced models like GPT-4.0, which is also restricted by a \$20 monthly subscription fee.

Given these issues, there's a growing interest in running foundation models for LLMs locally within enterprises, to avoid sending sensitive information to the cloud. However, training these models is expensive. GPT-3, with its 174 billion parameters, had an estimated training cost of \$1.8 million. Similarly, the European model Bloom had an estimated training cost of \$2.3 million, and GPT-4's estimated training cost is around \$63 million.

The ecological impact of training these models is also significant, mainly due to CO₂ emissions. For example, training GPT-3 in May 2020 was estimated to produce 502 tons of CO₂, equivalent to the complete lifetime emissions of nearly 8 cars, including the fuel consumed. However, there has been improvement in efficiency over time. Meta's OPT model, with 174 billion parameters, produced only 70 tons of CO₂, demonstrating a significant reduction (7x) in emissions in two years. In July 2022, the Bloom model, also with 176 billion parameters, only emitted 25 tons of CO₂, so 20x less.

In Europe, efforts have been made to develop models like Bloom, an open-source alternative to

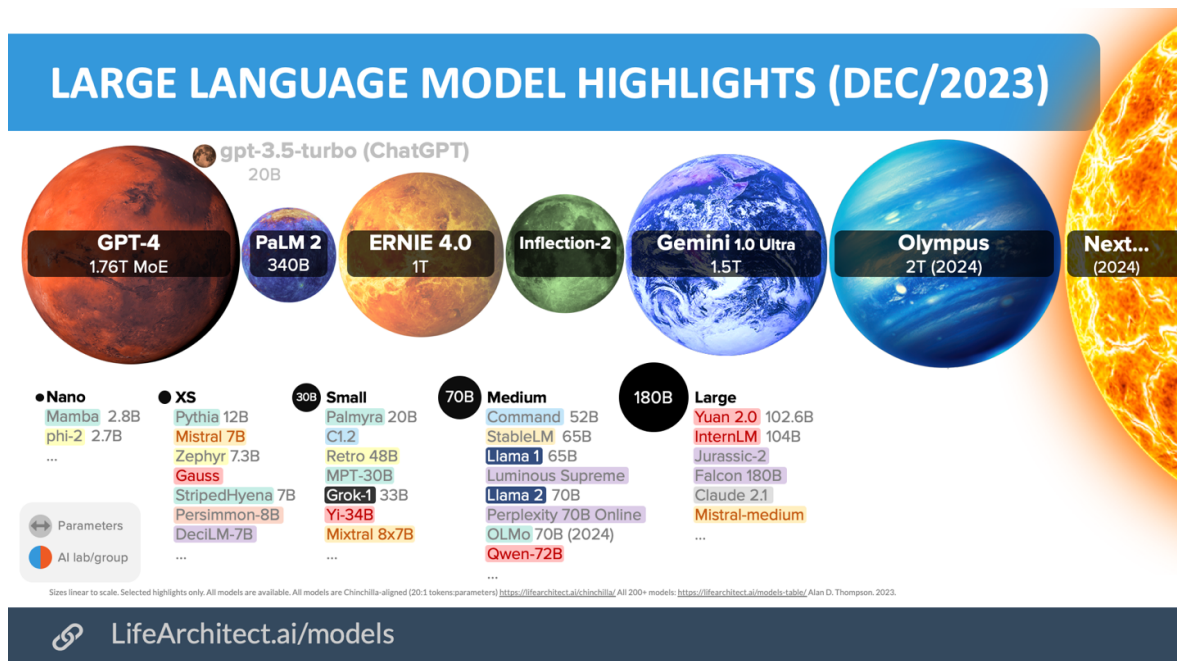


Figure 10: The always increasing size of foundation models. From [https:// life architect.ai/ models/](https://life architect.ai/models/)

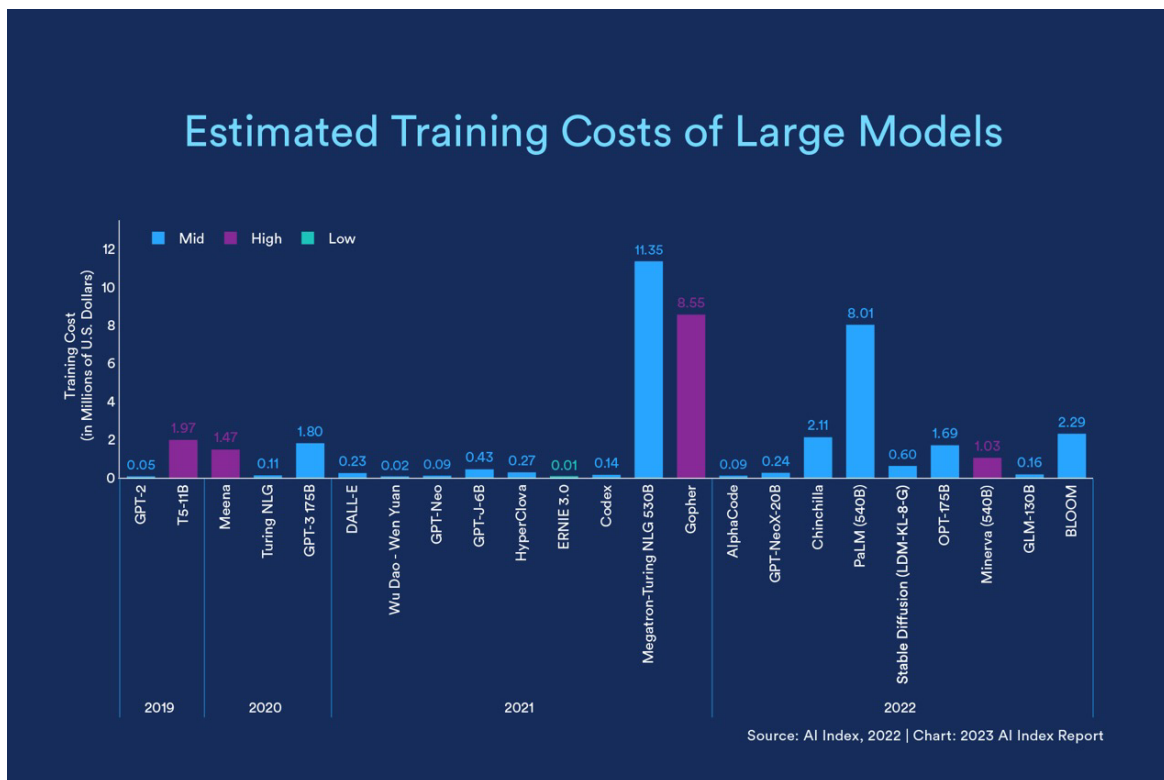


Figure 11: Training large foundation models is not cheap! From “2023 State of AI in 14 Charts” available at <https://hai.stanford.edu/news/2023-state-ai-14-charts>

GPT-3. During one-year, from May 2021 to May 2022, more than 1,000 researchers from 60 countries and more than 250 institutions are creating together a very large multilingual neural network language model and a very large multilingual text dataset on the 28 petaflops Jean Zay (IDRIS) supercomputer located near Paris, France, which mostly uses nuclear energy. The energy generated by Jean Zay is repurposed for heating, contributing to its low CO₂ footprint. Bloom was trained on 1.5 terabytes of text, it supports 43 languages and 16 programming languages, and took 118 days of training on 384 A100 GPUs. It also includes smaller versions tailored for instruction following, meaning it can answer questions and act as a chatbot¹⁰.

The democratization of large language models (LLMs) was significantly advanced with the release of Meta’s LLaMA foundation models in February 2023. These models ranged from 7 billion to 64 billion parameters. Initially available only to researchers, they were leaked online a week after their announcement, sparking a surge in the development of fine-tuned large models.

In March 2023, Stanford launched Alpaca[5], a robust, easily replicable instruction-following model based on the LLaMA 7B model, fine-tuned with 52K instruction-following demonstrations. This model was cost-effectively fine-tuned for under \$600, and its code was released on GitHub, further driving the development of an explosion of new models based on Meta’s leaked foundation model. The Figure 13 shows how LLaMA triggered a complete ecosystems of Open source models (in fact, we prefer to call them “open parameters” has the training database is not always provided - this is the case for LLaMA and LLaMA-2).

On July 18, 2023, Meta introduced LLaMA-2, a model free for both research and commercial use, easily accessible in repositories and HuggingFace, a California-based company founded by French entrepreneurs, emerged as a main repository for open LLMs. HuggingFace is akin to GitHub for the AI community, hosting over 450,000 models from universities and companies like Meta, but also Microsoft, Stability.ai, Mistral.ai and also OpenAI. The community and HuggingFace have

¹⁰ From <https://bigscience.huggingface.co/>

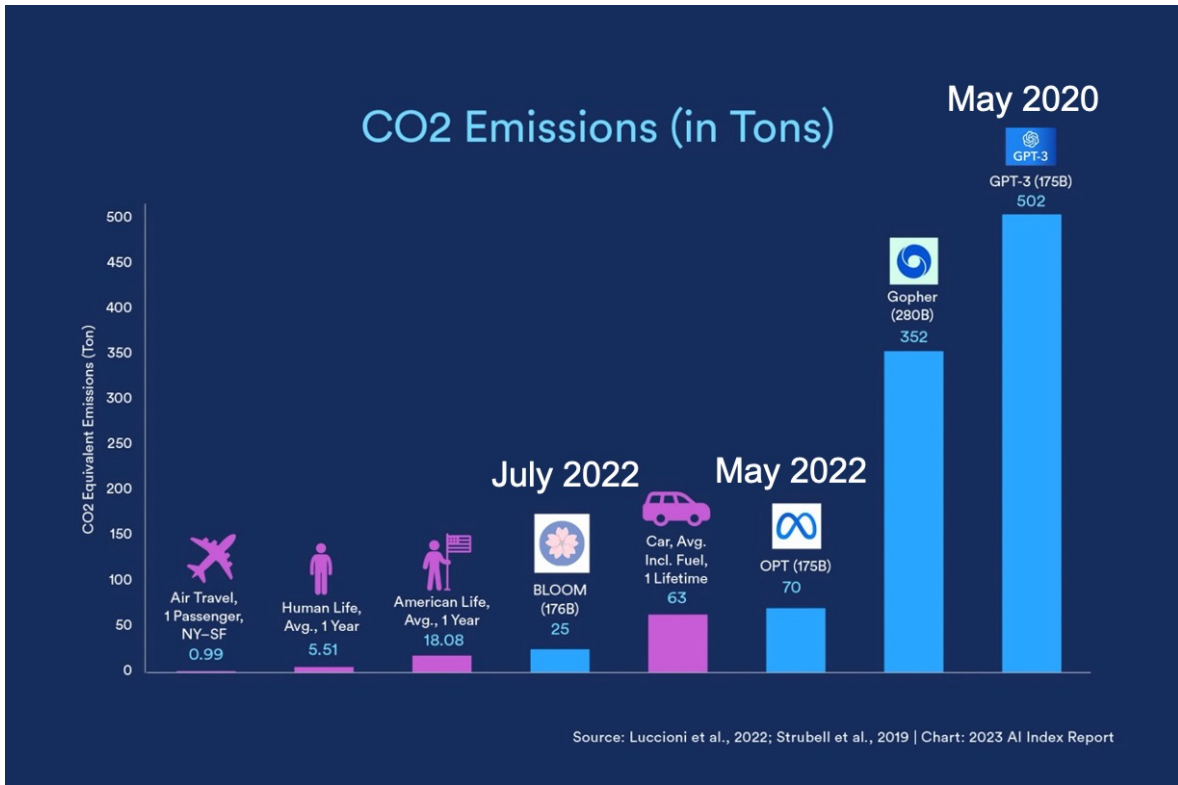


Figure 12: CO₂ impact for the creation of foundation models. From “2023 State of AI in 14 Charts” available at <https://hai.stanford.edu/news/2023-state-ai-14-charts>

made it possible to download and run models locally on personal computers. Most computers can run models up to 13 billion parameters. Running larger models requires GPUs with larger memory. For example, a 13 billion parameter LLaMA-2 model could run on a Mac Mini with comparable speed to ChatGPT while consuming only 21W.

The capability to run LLMs locally on smartphones has led to many developments. Companies like Meta are working (with Qualcomm) to optimize the execution of models like LLaMA-2 directly on smartphone chips. On-device AI offers several advantages, including cost reduction (no entrance fee or cloud costs), improved reliability (it is always available, it doesn't depend on the load of the servers) and performance, and enhanced privacy and security (everything is running locally, no information leaves the smartphone). It also allows for personalization,

like digital assistants that can be refined locally adapted to the user.

In October 2023, Qualcomm announced¹¹ its Snapdragon 8 Gen 3[7] with AI capabilities to perform multimodal generative AI models, including popular large language and vision models, transformer networks for speech recognition up to 10 billion parameters solely on the device. The Snapdragon 8 Gen 3 will be available in smartphones sold in 2024. The estimated power consumption of the SoC is lower than 10W (TbC).

MediaTek also announced¹² its Dimensity 9300 chip[8] in November 2023, capable of running a 7 billion parameter large language model at 20 tokens per second. It can also run a 13 GB model compressed to 5 GB to fit into RAM. The Dimensity 9300 supports NeuroPilot Fusion (which can

11 <https://www.qualcomm.com/products/mobile/snapdragon/smartphones/snapdragon-8-series-mobile-platforms/snapdragon-8-gen-3-mobile-platform>

12 From <https://www.mediatek.com/products/smartphones-2/mediatek-dimensity-9300>

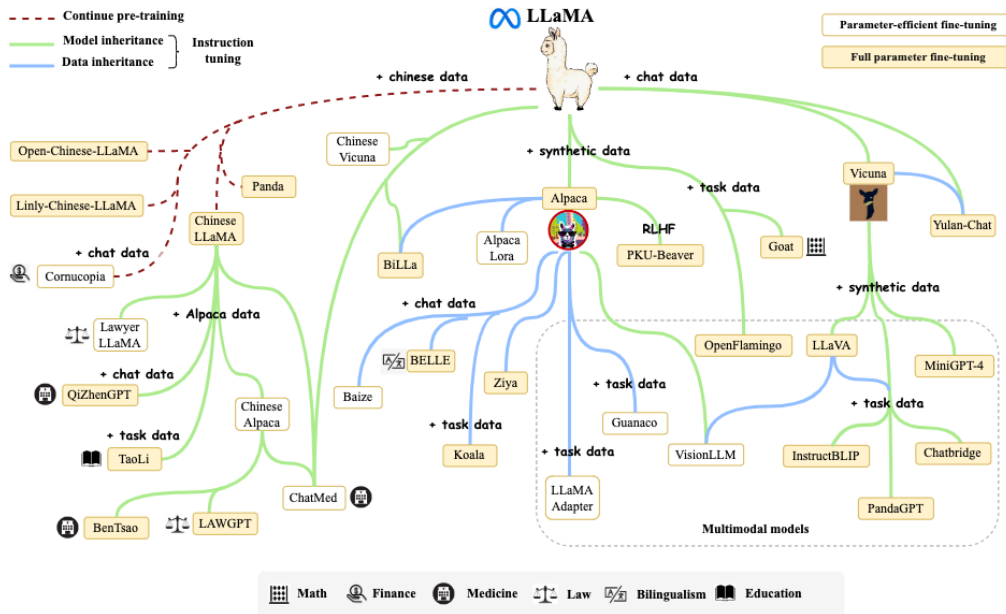


Figure 13 : An evolutionary graph of the research work conducted on LLaMA. Due to the huge number, we cannot include all the LLaMA variants in this figure, even much excellent work. From [6], also available at <https://github.com/RUCAIBox/LLMSurvey>

```

generate: n_ctx = 4096, n_batch = 512, n_predict = -1, n_keep = 1

== Running in interactive mode. ==
- Press Ctrl+C to interject at any time.
- Press Return to return control to LLaMa.
- To return control without starting a new line, end your input with '/'.
- If you want to submit another line, end your input with '\n'.

> Write a code that displays the 100 first prime numbers.
Here's an implementation in Python to display the first 100 prime numbers:
```python
primes = []
for num in range(2, 101):
 is_prime = True
 for i in range(2, int(num ** 0.5) + 1):
 if num % i == 0:
 is_prime = False
 break
 else:
 primes.append(num)
 continue
for prime in primes[:100]:
 print(prime)
```

This code uses a simple approach to find the first 100 prime numbers. It iterates through all numbers from 2 to 99 (since we don't consider 0 and 1 as prime numbers). For each number, it checks if it is divisible by any smaller integer using an inner loop. If the number remains undivided after checking all smaller integers up to its square root, then it is considered a prime number and added to our list of primes ('primes'). Finally, we print the first 100 elements from this list.
    
```

Figure 14: Example of a LLM (here SOLAR-10.7B-Instruct) running on a consumer grade PC (in this case a Mac mini) only consuming 21W while running.

continuously perform LoRA low-rank adaptation), enabling fine-tuning on the foundation model locally on the smartphone.

From these announcements, it is clear that generative AI, and AI in general will be the key selling factor of 2024 high-end smartphones, and that LLMs or models with up to 10B parameters will be able to run on smartphones, jeopardizing some of the market of the AI in the cloud.

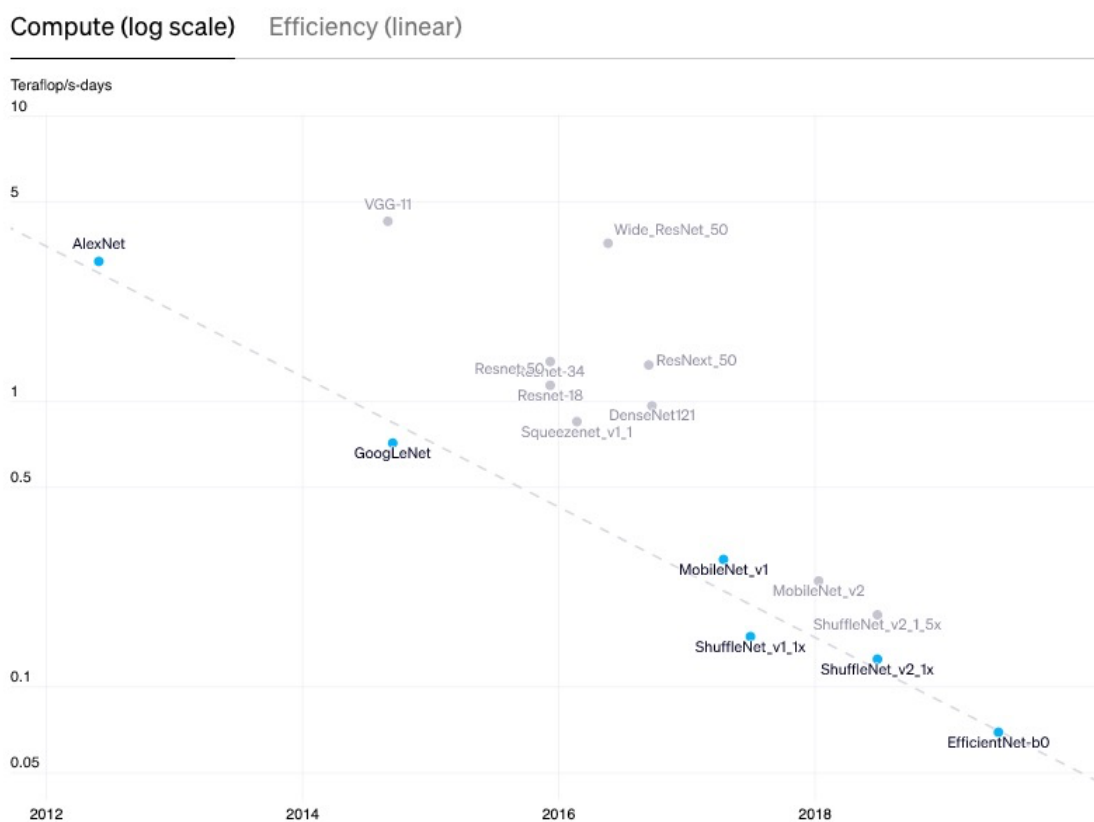
Smaller foundations models are more and more efficient, and can already run on devices like smartphones

Smaller LLMs, such as those with 7 or 10 billion parameters, may not perform at the same level as models with 174 billion parameters or more. However, there have been algorithmic improvements since 2022, leading to a decrease in the computational power needed to train a neural network to achieve similar performance. This reduction in necessary compute power has been observed to double approximately every 16 months, a trend akin to a law for neural networks. For instance, training GPT-3 in May 2020 required seven times more computational power than training Meta’s OPT model two years later in May 2022 for comparable performance.

An ultra simplified chronology of 2023 of “open-parameters” foundation models

February 2023:

- Meta AI introduced the LLaMA Model Series, designed for high performance within specific computing budgets.
- The strategy to create these models involved training smaller models on more data and for longer periods to enhance performance.
- The largest model in this series had 65 billion parameters trained on 1.4 trillion tokens. Other models had 6 billion and 13 billion parameters trained on 1 trillion tokens.
- The 13 billion parameter model outperformed GPT-3 in many benchmarks, and the largest model was a leading performer upon its release.



Total amount of compute in teraflops/s-days used to train to AlexNet level performance.

Figure 15: Total amount of compute in teraflops/s-days used to train to AlexNet level performance. 44x less compute required to get to AlexNet performance 7 years later. From <https://openai.com/research/ai-and-efficiency>

- These models were distributed under a non-commercial license.

July 2023:

- The LLaMA-2 family of models was released, ranging from 7 billion to 70 billion parameters.
- These models were trained on 2 trillion tokens from publicly available data.
- The fine-tuning process included Reinforcement Learning from Human Feedback (RLHF), known as the alignment procedure.
- They were free for research and commercial use.

September 2023:

- Mistral AI, a new Gen AI startup, released the Mistral-7B model.
- The model was trained on an undisclosed number of tokens extracted from the open web.
- Mistral-7B introduced new architecture choices, including Grouped-Query Attention, Sliding-window Attention, and Byte-fallback BPE tokenizer (that ensures that characters are never mapped to out of vocabulary tokens). The model was released in Base and Instruct Versions and licensed under Apache-2.0.

November 2023:

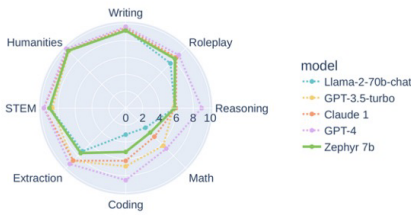
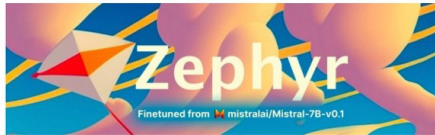
- O1-AI from China introduced the Yi Model family.
- The family included two model variations with 6 billion to 34 billion parameters.
- The models were bilingual, trained on both English and Chinese, using 3 trillion tokens.
- Yi-34/6B-200K supported up to a 200,000 context window and topped the Open LLM leaderboard and new benchmarks.

December 2023:

- Upstage AI from Korea released the SOLAR-10.7B model.
- The model, with approximately 11 billion parameters, used a new merging technique called "Depth Up-Scaling."
- Its architecture was based on Llama2 (7 billion with Mistral 7 billion weights), with two 7 billion models stacked together.
- The training data for this model, both open and proprietary, was undisclosed.

Similarly, smaller yet more powerful LLMs are becoming suited for smart devices. New models derived from LLaMA 7B or Mistral 7B are now freely available for download and offer benchmark results close to those of models with 25 times more parameters, like GPT-3.5. For example, Zephyr, a fine-tuned version of Mistral's 7B model, and OpenChat 3.5, a model fine-tuned using C-RLFT - C(onditioned)-RLFT(reinforcement learning fine-tuning) - with Mistral 7B as the base, both show impressive performance in some benchmarks.

The competition in the field is intense, with the goal being to create smaller, more efficient LLMs that maintain high performance. There are new techniques emerging constantly. For instance, after the success of the 'mixture of experts' approach (demonstrated by Mistral.ai's Mixtral 8x7B, a high-quality sparse mixture of experts model (SMoE) with open weights), technical developments like 'depth up-scaling (DUS)' have enabled more efficient LLMs. Depth up-scaling simplifies the upscaling process of LLMs, as demonstrated by models like SOLAR 10.7 billion parameter LLM (its architecture was based on Llama2 - 7 billion with Mistral 7 billion weights- , with two 7 billion models stacked together), which shows strong performance in various natural language processing tasks.



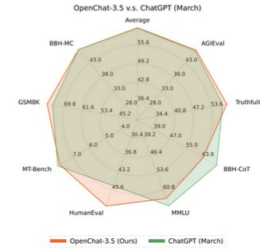
From <https://huggingface.co/HuggingFaceH4/zephyr-7b-beta>



[Online Demo](#) • [Discord](#) • [Huggingface](#) • [Paper](#)

🔥 The first 7B model Achieves Comparable Results with ChatGPT (March)! 🔥

🏆 #1 Open-source model on MT-bench scoring 7.81, outperforming 70B models 🏆



From <https://github.com/imoneoi/openchat>

Figure 16: Some "small" models with good results

Multiple optimizations are also contributing to the ability to run LLMs and transformer-based neural networks on devices with lower power and performance. For example, as evidenced in the Figure 18, optimizing a model to run on a CPU with efficient data management can improve efficiency by a factor of 6.7. Transitioning to GPU and leveraging specific floating-point operations further enhances performance. Additionally, optimizing transformer architecture can lead to even greater performance gains. Overall, by optimizing both platforms and algorithms, a significant improvement of factor 810x (nearly 4 orders of magnitude!) can be achieved.

| | LLaMA 2 70B | GPT - 3.5 | Mixtral 8x7B |
|--|--------------|-------------|--------------|
| MMLU
(MCQ in 57 subjects) | 69.9% | 70.0% | 70.6% |
| HellaSwag
(10-shot) | 87.1% | 85.5% | 86.7% |
| ARC Challenge
(25-shot) | 85.1% | 85.2% | 85.8% |
| WinoGrande
(5-shot) | 83.2% | 81.6% | 81.2% |
| MBPP
(pass@1) | 49.8% | 52.2% | 60.7% |
| GSM-8K
(5-shot) | 53.6% | 57.1% | 58.4% |
| MT Bench
(for Instruct Models) | 6.86 | 8.32 | 8.30 |

Figure 17 : Mixtral 8x7B, a high-quality sparse mixture of experts model (SMoE) was released by Mistral.ai in December 2023. Mixtral outperforms LLaMA 2 70B on most benchmarks with 6x faster inference. In particular, it matches or outperforms GPT3.5 on most standard benchmarks. From <https://mistral.ai/news/mixtral-of-experts/>

¹³ On some benchmarks. However, several specialized models can work together, therefore enlarging the global capabilities of this set of models. This trend of having a set of several more specialized

With the development of SoC able to execute (and even fine tune) models up to 10s of Billions of parameters, the fact that models of this size can be fine-tuned or architected to be as performing as larger models¹³, it is clear that we are entering in the era of the "continuum of computing", where some AI tasks will be done locally at the edge, and others still in the cloud. With the progress in model design and on hardware accelerators, (generative) artificial intelligence will run on always "smaller" devices in the future.

Carbon Optimization via HW-SW Co-Design

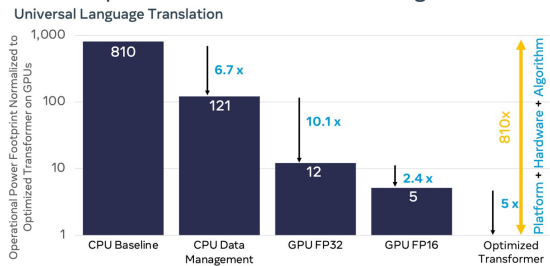


Figure 18: Carole-Jean Wu, Meta, "Designing computer systems for sustainability", ACACES, Italy, July 2023.

models working together is in line with the ideas of the Next Computing Paradigm, where a federation of smart agent are working collectively to achieve a goal that one agent cannot achieve alone.

Conclusion and recommendations

The launch of ChatGPT in November 2022 marked a significant milestone in the world of technology, creating a substantial impact. The recent advancements in AI, particularly in generative AI based on transformer technology, heavily depend on hardware technologies. Progress in the foundation models, new models architectures enable smaller models to be efficient for practical tasks. Technology is being propelled forward by the development of new accelerators, which enable smaller models to operate on edge devices.

Looking ahead, the evolution of language models is poised to take an exciting turn. Future language models are expected to move beyond learning words and texts to embracing multi-modality, learning through images and other modalities. This evolution will likely involve a combination of more specialized models with a 'mixture of experts' approach, moving away from monolithic models towards collaborative edge systems. These advanced models are also being linked to digital twins to experience and understand the laws of physics, as exemplified by applications like Isaac Sim¹⁴. This proactive approach allows models to directly interact with and learn from the world, potentially leading to their integration into robots and other autonomous systems.

Therefore, it is clear that Europe should take a leading role in the development of foundational AI models, with a focus on aligning them with "European" values. This involves creating and disseminating methodologies and datasets specifically tailored to meet regional requirements, ensuring AI sovereignty and promoting a digital economy reflective of European standards and ethics. In addition, there's an emphasis on promoting open-source models to foster a collaborative AI ecosystem with open access to AI resources. This approach nurtures a culture of shared progress. Furthermore, Europe aims to spearhead the integration of Large Language Models (LLMs) into smart devices, focusing on local solutions and specialized accelerators. This strategy will empower real-time AI applications on edge devices, reduce reliance on centralized data centers, enhance privacy, and increase efficiency. The plan includes supporting both startups and established companies to develop edge AI capabilities, thereby building a decentralized and resilient AI infrastructure.

¹⁴ <https://developer.nvidia.com/isaac-sim>

References

- [1] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", Advances in Neural Information Processing Systems, F. Pereira and C.J. Burges and L. Bottou and K.Q. Weinberger editors, Vol. 25, 2012, [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf. [Accessed 14 December 2023].
- [2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, "Attention Is All You Need", [Online]. <https://arxiv.org/abs/1706.03762>. [Accessed 29 November 2023].
- [3] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, "LoRA: Low-Rank Adaptation of Large Language Models", [Online]. <https://arxiv.org/abs/2106.09685>. [Accessed 29 November 2023].
- [4] Albert Reuther, Peter Michaleas, Michael Jones, Vijay Gadepally, Siddharth Samsi, Jeremy Kepner, "AI and ML Accelerator Survey and Trends", [Online]. Available: <https://arxiv.org/abs/2210.04055>. [Accessed 29 November 2023].
- [5] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, Tatsunori B. Hashimoto, "Alpaca: A Strong, Replicable Instruction-Following Model", Stanford University, [Online]. Available: <https://crfm.stanford.edu/2023/03/13/alpaca.html>. [Accessed 29 November 2023].
- [6] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, Ji-Rong Wen, "A Survey of Large Language Models", [Online]. <https://arxiv.org/abs/2303.18223>. [Accessed 29 November 2023].
- [7] "Snapdragon 8 Gen 3 Mobile Platform," [Online]. Available: <https://www.qualcomm.com/products/mobile/snapdragon/smartphones/snapdragon-8-series-mobile-platforms/snapdragon-8-gen-3-mobile-platform>. [Accessed 14 December 2023].
- [8] "MediaTek Dimensity 9300 deep dive: A true Snapdragon rival?," 6 November 2023. [Online]. Available: <https://www.androidauthority.com/mediatek-dimensity-9300-explained-3381678/>. [Accessed 14 December 2023].

ChatGPT is a Chat bot from OpenAI using the GPT-4 Large Language Model technology.

Marc Duranton is a researcher in the research and technology department at CEA (the French Atomic Energy Commission) and the coordinator of the HiPEAC Vision 2024.



AI will radically change how we do software engineering in the future.

AI-Assisted Software Engineering (AISE)

by **Andreas Metzger (Editor)**

Generative AI, large language models and AI chatbots are new tools that can help humans with various creative activities. One of these activities is software engineering, which will be radically transformed by the use of AI. We call this AI-Assisted Software Engineering (AISE). AISE will have an impact on all parts of software engineering. We explore the opportunities but also important challenges of AISE. We discuss how AISE will affect the different software engineering tasks, and discuss important cross-cutting concerns, such as the software life-cycle, software security, as well as non-technical concerns.

Key insights

- AISE will enable software developers to focus more on the problem and the quality, and less on the details and the programmatic syntax. It will significantly increase the level of abstraction, making software engineering much more automated and productive.
- AISE “*has the potential to change the software profession more than any other recent technology*”³.
- AISE will help the European software industry to compete globally and deliver high-quality software.

Key recommendations

- Delivering AISE’s full potential requires addressing significant technical and non-technical challenges. Addressing these challenges requires dedicated research and innovation (R&I) actions to facilitate the competitiveness of the European primary and secondary software industry.
- R&I actions on AISE will be key to addressing gaps identified in the Horizon Europe Strategic Plan 2025-2027 Analysisⁱⁱ, which states that “Regarding AI for Software Engineering, ‘software co-engineering’, the following aspects of the software development lifecycle are not covered yet: code search using natural language processing techniques, using AI to analyse the code quality, perform automatic bug fixing, or the application of AI-based failure prediction algorithms for the software at operation time, have, at this stage not been fully incorporated; low-code approaches will increase software quality and productivity, wider take-up, and ease of deployment, also from close-to-the-user configuration and personalisation.”

AI-Assisted Requirements Engineering

Requirements engineering is the process of defining and refining the needs and expectations of different stakeholders for a software system. Requirements engineering helps to establish a common understanding and agreement on the requirements among all the parties involved in the software engineering process. Requirements engineering uses various methods and techniques to support the whole requirements life cycle, from gathering and analysing requirements to validating and managing themⁱⁱⁱ.

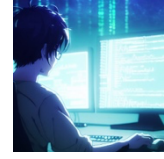


Opportunities. AI combined with natural language processing can assist requirements engineering activities, since many requirements are expressed in natural language^{iv}. In particular AI Chatbots based on LLMs can deliver untapped automation opportunities in requirements engineering. AI chatbots can improve the communication and understanding with people from different backgrounds and skill levels, as well as non-technical users. Moreover, it may help to overcome the language barriers; for example, allowing the stakeholders to converse and provide input and feedback in their preferred language.

Challenges. AI chatbots are fascinating because they can answer natural language questions and generate rich text, but they also pose a problem for requirements engineering. The big language model behind an AI chatbot might "hallucinate", i.e., create nonsensical text that does not match the input. This means AI chatbots might have low fidelity. Also, AI chatbots might give different answers to the same question, which means they have low stability. Low fidelity and stability can lower the quality of the responses and affect the requirements engineering outcomes, leading to poor requirements. A key challenge is how to make sure that AI chatbots produce high-fidelity and high-stability responses. This might involve using specific prompt engineering that takes into account the requirements engineering context, as well as fine-tuning the parameters of the underlying LLMs.

AI-Assisted Code Creation and Maintenance

Code creation and maintenance are difficult because they require designing effective algorithms and programmatic solutions, ensuring compatibility with different platforms, and fixing errors or bugs that may arise. As software becomes more complex and sophisticated, it also becomes more challenging to ensure its quality, performance, and flexibility.



Generative AI tools (such as GitHub Copilot and Code Llama) are transforming programming, offering profound improvements in code creation, maintenance, and thereby significantly increase developer productivity. The adoption of such AI tools in SE is estimated to lead to a \$1.5 trillion boost in global GDP by 2030^v.

Opportunities. Generative AI tools make code suggestions and autocompletions smarter, providing developers with code snippets that they may not have thought of. Programming is being supplemented by prompt engineering to generate code from natural language prompts. Generative AI tools become increasingly interactive (such as GPT Engineer) leading programmers to refine and improve their prompts to get the desired results, ultimately enhancing computational thinking and change how coding skills are learned.

Generative AI can make code refactoring and restructuring easier and faster, enhancing performance, readability, and modularity over time. Generative AI can also quickly generate patches and fixes for vulnerabilities and bugs, reducing the time to address critical issues. Engineers can use conversational interfaces to fine-tune and tailor AI-generated adaptations to their specific context, rather than accepting generic suggestions.

Generative AI can also help with legacy code maintenance, which is becoming more challenging as the expertise of legacy programming languages and environments fades away. Generative AI tools can analyse legacy code, find quality problems, and suggest refactoring or rewriting options (such as IBM's watsonX Code Assistant that transforms COBOL code into Java).

Challenges. AISE offers novel possibilities for code generation and software engineering, but it also faces important difficulties. For example, LLMs can produce outputs that are not consistent

or logical (hallucinations, as mentioned above), which can affect the quality and traceability of the code. Moreover, security is a major concern, as tools such as GitHub Copilot may generate code with vulnerabilities in about 40% of the cases^{vi}. Also, intellectual property issues can also emerge, as some generated code may unintentionally copy licensed programs (discussed further below, under “Non-Technical Concerns in AISE”). While Generative AI can produce high-quality code for general purpose situations, generating domain-specific code may still be limited. Here, retrieval-augmented generation can enhance code generation and summarization by seamlessly pulling relevant coding patterns from relevant databases, and aiding maintenance by ensuring consistent and up-to-date coding practices.

AI-Assisted Quality Assurance and Testing

As our dependency on software continues to grow, we see an increasing pressure on software engineering to quickly deliver software while assuring adequate software quality. The total cost of poor software quality in the US^{vii} alone amounted to 2.08 trillion USD in 2020. Given this central and crucial role of quality assurance, the use of AI for quality assurance has been widely studied in research. Even before the launch of ChatGPT (in November 2022), research delivered over 800 publications^{viii} with a compound annual growth rate of 38% (between 2018 and 2022).



Opportunities. Generative AI, LLMs and AI Chatbots will help evaluate code quality and make the quality process more efficient. Some examples are grouping and ranking tests and finding parts of the code that are more likely to have bugs. AISE will offer novel possibilities for quality assurance, such as generating better test assertions, more precisely fixing programs, and more accurately pinpointing bugs introduced during software changes.

Challenges. When using Generative AI, the generated tests may contain bugs in the test inputs or the expected outcomes. Generating the expected outcomes for the tests is difficult (this is known as the test oracle problem), as it may require information that is beyond the current capabilities of AI. Moreover, LLMs may produce unrealistic or erroneous outputs, such as calling non-existent

functions or passing unsupported parameters to existing functions. Tests that do not have any compile or runtime errors still need to be checked for correctness. Additionally, when the tests fail, it is hard to determine whether the fault lies in the software itself or in the test. This calls for a systematic approach of how AI-generated tests are incorporated into the quality assurance process.

AI can help predict faults, failures and development effort in novel ways. These predictions can assist practitioners in making timely decisions. However, AI models may sometimes produce inaccurate predictions, which can result in wasting resources on false positives or overlooking important actions due to false negatives. Providing AI models with ways to express and guarantee their confidence in their outcomes would greatly improve the quality of AI-supported decision making.

AI-Assisted Integration and Deployment

To deliver software changes faster and more frequently, modern software development practices, such as DevOps, use automated Continuous Integration and Deployment (CI/CD) pipelines. This facilitates the rapid release and operation of new software versions.



Opportunities. Writing scripts for CI/CD pipelines can be complex and tedious for human developers and thus presents itself as a promising target for Generative AI. This can be combined with novel “virtualization” techniques to hide the low-level details of computing, storage and networking resources, such as Function as a Service (FaaS) or Serverless Architecture, where developers only specify the resource requirements for their software, without worrying about the deployment targets.

As the computing infrastructure shifts from centralized cloud to decentralized Cloud-Edge-IoT continuum, AI becomes an essential tool for managing and supporting the deployment of software on the available resources. To handle the ever-expanding computing continuum for continuous delivery, advanced learning and reasoning capabilities (such as multi-agent reinforcement learning, deep reinforcement learning, and “meta” reinforcement learning) offer novel opportunities to

capture the dynamicity, complexity and uncertainty of the Cloud-Edge-IoT continuum.

Challenges. AI can help automate the integration and deployment of software, but it faces many challenges, including (1) How to plan the deployment with multiple objectives, preferences, and constraints that may vary depending on the situation; (2) How to capture the high-level intents of developers instead of low-level actions that may not reflect their goals; (3) How to keep developers in the loop and let them review and refine the deployment scripts over time; (4) How to coordinate the deployment agents across different locations, domains, or providers to achieve a global optimal solution; (5) How to build trust between developers and AI by explaining the deployment decisions and outcomes.

AI-Assisted Adaptation

A software system that can change its own structure and behaviour at runtime is called a (self-)adaptive software system. Examples of adaptive software systems are cloud systems that can scale up or down, IoT systems that can learn and act intelligently, and process management systems that can anticipate and adjust to changes.



A crucial component of an adaptive system is its adaptation logic, which specifies when and how the system should adapt itself. When developing the adaptation logic, developers have to deal with design time uncertainty, which means they have incomplete information about when and how the system should adapt^{ix}.

Opportunities. Using AI in an online fashion (e.g., online deep learning) is a new way to deal with design time uncertainty for adaptive systems. For example, by using deep reinforcement learning at run time, the system can learn from data that is only available when it is running, and thus cope with uncertainty better. Moreover, the system can benefit from predictive monitoring, which is a technique that estimates the likelihood of failures in the near future by leveraging deep learning models and complex operational data.

Challenges. Online reinforcement learning has been used successfully for adaptive systems, but important challenges remain. Modern deep learning algorithms are stochastic, which means that

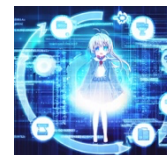
their performance can significantly vary. This means that even if deep learning models perform well in the lab, they may not perform well in an actual production environment. While factors that cause variance may be controlled (e.g., non-deterministic neural network layers or random weight initialization), this typically leads to performance degradation. A more promising way thus is to use "meta"-learning techniques on top of the deep learning models, such as ensemble learning, which aggregates and thereby enhances the results of multiple deep learning models.

The successful application of online reinforcement learning depends on how well the learning problem, and in particular the reward function, is defined. Software engineers need to explicitly define a reward function, which quantifies the feedback to the RL algorithm. Getting the reward function right such that it accurately reflects the trade-off among different goal dimensions and achieves its overall learning goal is a challenge.

In addition to leveraging AI chatbots to provide natural-language and interactive explanations and thereby support debugging of adaptive systems^x, a novel direction may be to use LLMs to generate suitable reward functions from a sufficiently concise description of the problem context and overall learning goal.

AI-Assisted Software Life-Cycle

As highlighted above, AISE will transform software engineering by enabling higher levels of automation for various tasks, from requirements engineering to operations and legacy systems maintenance. This will improve the efficiency and speed of the software life cycle and supply chain. Many solutions have been proposed to automate specific software engineering tasks using AI techniques. A recent survey^{xi} indicates that the research output on using ML for automating individual software engineering tasks shows a compound annual growth rate of 33% (from 2018 to 2020).



Opportunities. Automating individual software engineering tasks is beneficial, but there is also a great potential to exploit synergies among these tasks across the whole software life cycle and supply chain. For instance, a single quality assurance task may not be enough to ensure the desired software quality. Ideally, one would use a suitable

combination of different tasks – such as dynamic testing and static code analysis. Also, exploiting the synergies between different tasks can increase the effectiveness of the individual tasks. For example, having a good estimate of the fault density of a software component, e.g., by using deep learning-based fault prediction methods, could be used to optimise and prioritise testing effort and budget.

Challenges. One way to leverage synergies in AISE is to use the output of one AI-based software engineering task as input for another task. For example, one can use the predicted fault-proneness of a component to prioritize test cases. However, this approach may not fully exploit the potential of AI. A more interesting direction is to leverage synergies that take into account the specific characteristics of AI models. For example, an AI model can provide an explanation that makes the output of deep learning models more understandable, and thus more trustworthy, when used as input for another activity.

AI-Assisted Security

The software supply chain consists of many elements, such as code, tools, people, and processes. These elements interact with each other to create, test, deploy, and maintain software products. However, this also exposes the software supply chain to various threats (e.g., as discussed in the NESSI position paper on software security^{xiii}). To protect the software supply chain from cyber-attacks, which are becoming more frequent and advanced, AI can be a useful tool; for example, the US White House has launched the AI Cyber Challenge to Protect America’s Critical Software.



Opportunities. AI can enhance efficiency, productivity, and quality in software engineering and security, as discussed in previous sections. AI can help programmers develop secure code by suggesting secure code snippets. AI can generate more comprehensive test cases that improve test coverage. AI can improve patch management by detecting vulnerabilities and by predicting the impact of a patch on the system’s stability before deployment. Other promising opportunities include analysing large amounts of data for better insights and contextual awareness (e.g., anomaly detection), learning and adapting to changing conditions and predicting new threats, and drawing

conclusions and generating action plans in real-time (e.g., incident response).

AI has the potential to improve security in all areas of the software supply chain and life cycle; e.g., via software verification and validation, software composition analysis, certification and conformity assessment, risk management for software-intensive complex systems, vulnerability detection, maintenance, reporting, and threat intelligence sharing.

Challenges. AI for secure software engineering still faces many shortcomings including generation of insecure code, uncertain prediction quality, inadequate reasoning, and low transparency. High-quality datasets are crucial for training secure AI models. For example, curated datasets may be needed to train and to validate AI models so they generate more secure code with less vulnerabilities. To ensure that software engineers can comprehend, modify, and update the code, AI-assisted code generation should be transparent and explainable.

Unfortunately, AI tools for software security may also be misused by hackers and cyber-criminals. Software developers and security professionals should understand the features and impacts of AI-based cyber-attacks and how to stop them. The AI tools themselves may be vulnerable and need security measures to reduce their AI-specific security threats.

AI adoption for security is limited due to integration difficulties, a shortage of skilled professionals, and concerns about accuracy, dependability, and implementation costs. AI models, such as LLMs, pose security and privacy risks themselves, such as model inversion attacks revealing training data, bypassing safety features and using prompt injection for harmful outputs, data poisoning causing incorrect behaviour.

Non-Technical Concerns in AISE

Besides the technical concerns discussed above, AISE also offers novel non-technical opportunities but at the same time also introduces major non-technical challenges.



Opportunities. Generative AI can be used for various purposes that can improve innovation and

efficiency. It can teach developers new coding skills and languages by giving them interactive suggestions^{xiii}. It can also help developers brainstorm new ways of solving problems by generating code examples^{xiv}. Moreover, it can help developers evaluate the possible outcomes of their software, such as how it might affect different stakeholders, the company's ROI, and other ethical issues. Lastly, it can also help developers understand the legal aspects of software engineering, such as AI, data, and platform laws, to ensure compliance and avoid IPR violations.

In particular, AI chatbots based on LLMs can provide natural-language and interactive explanations for software systems that have AI and ML components. Explainability can help software and service providers comply with the relevant laws and regulations, such as the GDPR and the AI Act in the EU. Furthermore, explainability can help software and service users trust the software systems by understanding how they produce their results and whether they are acceptable or not.

Challenges.

Intellectual Property. AI-generated content, such as synthetic code, poses new legal questions for the current IP regimes that are based on human creativity^{xv}. How to give fair credit and licenses for the products of proprietary models with different levels of licensing is a controversial issue, especially when the models are trained on user-generated data that is taken from the internet without compensating the users. It is not clear whether AI systems that create content independently without human input can be regarded as authors or not. Therefore, there is a demand for more transparency and consistency in the ownership and licensing of AI-generated content, as AISE becomes more common. To safeguard the rights and interests of both AISE tool developers and users, new licensing schemes and attribution standards should be created and harmonised. To deal with the legal challenges caused by AI-generated code that does not involve significant human input, copyright law and liability models should be adjusted accordingly.

Transparency. LLMs are complex AI systems that do not reveal their inner workings easily, making it hard to apply traditional methods of technology governance on such non-transparent LLMs. Some methods have been suggested to explain parts of LLMs, but they still face many shortcomings, such as ensuring fidelity of explanations and the

difficulty of evaluating and comparing explanations. To ensure quality and security, companies could establish internal oversight boards that review the used AISE tools and underlying LLMs before they are deployed. These boards could evaluate the potential risks of biases, security breaches, and misalignment with engineering objectives. Additionally, random audits of the code generated by LLMs could help verify its stability.

Sustainability. LLMs are powerful tools but also very costly in terms of energy and resources. They can produce thousands of tonnes of CO₂ equivalent during training, which may increase as models get bigger^{xvi}. In addition, energy consumption during inference (i.e., when generating the outcomes) becomes a concern with the increased use and adoption of these AI models. To reduce the environmental impact of LLMs, potential directions may be to use green computing infrastructure (including processors) and to reuse and leverage existing models. We can also prune large models to make them smaller and faster without losing quality^{xvii}. While LLMs may offer benefits for sustainability, such as automating quality assurance and reducing human effort, we need more research to understand the trade-offs and best practices for LLMs across their lifecycle.

Overreliance. The code-generation capabilities of modern coding assistants are impressive. However, the quality of the underlying LLMs and thus their generated code may deteriorate over time, requiring proper tracking and testing in order not to rely on low-quality models. LLMs should help (hence the term AI-assisted), not replace, human and collective programming knowledge. Otherwise, the risks of wrong outputs and unethical design may increase due to less human supervision and expertise. To ensure that LLMs are aligned with ethical values and social norms, they should be developed and deployed with the participation of relevant stakeholders, including the public and the affected groups. Moreover, LLMs should be subject to constitutional principles and expert oversight, as well as feedback mechanisms that can improve their performance and reduce their harms. Educating people about the benefits and risks of LLMs is essential for fostering trust and awareness and preventing overreliance on these powerful AI models.

Andreas Metzger is Vice Chair of NESSI (the European Networked Software and Services Initiative – <https://nessi.eu/>), Professor of software engineering at the University of Duisburg-Essen, and Head of “Adaptive Systems” at paluno (the Ruhr Institute for Software Technology – <https://paluno.uni-due.de/en/>).

References

- i Christof Ebert, Panos Louridas: Generative AI for Software Practitioners. *IEEE Softw.* 40(4): 30-38, 2023
- ii European Commission, Directorate-General for Research and Innovation, Horizon Europe strategic plan 2025-2027 analysis, Publications Office of the European Union, 2023, <https://data.europa.eu/doi/10.2777/637816>
- iii Klaus Pohl. *Requirements Engineering: Fundamentals, Principles, and Techniques*, Springer, 2010
- iv Walid Maalej: From RSSE to BotSE: Potentials and Challenges Revisited after 15 Years. *arXiv preprint*, 2023; <https://arxiv.org/abs/2304.09308>
- v Thomas Dohmke, Marco Iansiti, and Greg Richards. "Sea Change in Software Development: Economic and Productivity Analysis of the AI-Powered Developer Lifecycle." *arXiv preprint*, 2023; <https://arxiv.org/abs/2306.15033>
- vi Pearce, Hammond, et al. "Asleep at the keyboard? assessing the security of github copilot's code contributions" *IEEE Symposium on Security and Privacy*, 2022
- vii Herb Krasner. *The cost of poor software quality in the US: a 2020 report*. Consortium for Information and Software Quality (CISQ), 2021
- viii Andreas Metzger; Xhulja Shahini; Johannes Haerkötter, Klaus Pohl. *A Systematic Literature Review of Machine Learning for Uncovering Software Faults and Failures*, 2023; <https://doi.org/10.5281/zenodo.7615631>
- ix Andreas Metzger, Clement Quinton, Zoltan Mann, Luciano Baresi, Klaus Pohl, “Realizing Self-Adaptive Systems via Online Reinforcement Learning and Feature-Model-guided Exploration”, *Springer Computing*, March, 2022, pp. 1–22, 2022; <https://doi.org/10.1007/s00607-022-01052-x>
- x Andreas Metzger, Jone Bartel, Jan Laufer, „An AI Chatbot for Explaining Deep Reinforcement Learning Decisions of Service-oriented Systems”, 21st Int'l Conference on Service-Oriented Computing (ICSOC 2023), 2023, Springer, 2023; <https://arxiv.org/abs/2309.14391>
- xi Simin Wang et al. *Machine/Deep Learning for Software Engineering: A Systematic Literature Review*. *IEEE Transactions on Software Engineering*, 49(3): 1188-1231, 2023
- xii <https://nessi.eu/nessi-paper-on-software-security/>
- xiii Steven I. Ross, Fernando Martinez, Stephanie Houde, Michael J. Muller, Justin D. Weisz, The programmer's assistant: Conversational interaction with a large language model for software development. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 2023
- xiv David Noever, Kevin Williams, Chatbots as Fluent Polyglots: Revisiting Breakthrough Code Snippets. *arXiv preprint*, 2023; <https://arxiv.org/abs/2301.03373>
- xv Giorgio Franceschelli, Mirco Musolesi, Copyright in generative deep learning. *Data & Policy*, 4, e17, 2022
- xvi <https://aiindex.stanford.edu/report/>
- xvii Elias Frantar, Dan Alistarh: SparseGPT: Massive Language Models Can be Accurately Pruned in One-Shot. *ICML 2023*: 10323-10337



Challenges in EDA: from operational research techniques to Artificial Intelligence strategies for chip design

by **Alix Munier Kordon** and **Lilia Zaourar**

Artificial Intelligence (AI) could be pivotal in reducing complex EDA flow by intelligently narrowing the exploration space in complement to combinatorial optimization approaches. One of its notable advantages is that it does not demand a high level of user tool expertise, making it accessible and user-friendly. This democratization of expertise ensures that individuals without specialized knowledge can benefit from AI-driven solutions. Moreover, AI could improve efficiency by reducing turnaround time through effective iterations and parallel runs. This acceleration in the workflow enhances productivity and allows for more thorough exploration and optimization, ultimately leading to more informed decision-making.

Key insights

- Nowadays, chips are present everywhere and anywhere in our daily life ranging from various computing facilities, to interconnected objects. This phenomenon poses significant challenges in designing chips that need to be more and more sophisticated.
- Managing and optimizing a myriad of possibilities and configurations involved in designing integrated circuits induce a combinatorial explosion in chip design due to the exponential increase in complexity as the number of elements and interactions grows. Addressing this complexity requires sophisticated tools

and algorithms to navigate the vast design space, optimize performance, and ensure efficient chip functionality.

- There is a demand and need for automated and high-performance tools to speed up the design process of chips while ensuring their quality.
- AI inspired techniques in chip design could offers a more dynamic data-driven decision efficient approaches in addition to usual operational research and combinatorial methods. Thus contributing to advancements in performance, power efficiency, and overall design quality.

Key recommendations

- Shift towards hybridization between AI and combinatorial optimization strategies to preserve and maintain explainability while taking advantage of available data in electronic design automation flow. This hybridization aims to combine the strengths of AI, which excels in handling complex patterns and large datasets, with combinatorial optimization strategies known for their transparency and interpretability. By striking this balance, the goal is to harness the benefits of AI's data-driven capabilities while ensuring that the decision-making process remains understandable and explainable, a crucial factor in chip design.

- Facilitating shared data exchange between academic researchers and industry would foster collaboration in developing advanced methods tailored to address diverse challenges encountered during the EDA design phases. This cooperative approach allows academic and industrial entities to leverage their expertise and resources, collectively contributing to creating high-performance solutions for various design problems. By sharing common data, the collaborative ecosystem benefits from a rich pool of information and experiences, ultimately enhancing the efficiency and effectiveness of methods applied in the design processes.
- Tools Interoperability in EDA flow is indispensable for achieving a cohesive, efficient, collaborative chip design process. It enables designers to harness the strengths of diverse tools while ensuring a seamless exchange of information, flexibility and tool diversity, holistic design exploration, and efficient data exchange, ultimately contributing to developing high-quality semiconductor products.
- The shift towards multi-criteria considerations in chip design entails expanding the range of factors beyond traditional metrics such as surface area, power consumption, and time. This evolution involves incorporating additional criteria, including but not limited to the environmental impact, such as carbon footprint. This broader perspective reflects a growing awareness of the need to assess and optimize semiconductor designs for performance and efficiency, with a focus on sustainability and environmental considerations.

Introduction

Moore's Law¹ has been a guiding principle for the semiconductor industry for several decades. Even if it is an empirical trend based on historical observations, it has dictated the evolution of semiconductors and has been pushing the community to the extreme. This "law" forms the cornerstone of the computer industry, greatly impacting the revolution in chip design. As a result, the complexity of semiconductor chips follows a trend of

almost doubling each year, adding substantial intricacy to their design.

Recent high-end processor chips integrate more than 100 billion transistors, and Cerebras even integrates 2.6 Trillion transistors on their 850,000 cores Wafer-Scale Engine.

Over time, sustaining this rate of transistor doubling has become increasingly challenging due to physical and technological limitations and leading to a considerable combinatorial explosion. Therefore, as semiconductor technology advances, it becomes more difficult and costly to continue the same rate of transistor density increase. It has led to innovations such as 3D stacking, new materials, and alternative computing architectures to extend the capabilities of microchips.

How about the complexity battle?

Meanwhile, as technology nodes advance and application demands become more intricate, a concomitant increase in constraints becomes apparent. These constraints encompass a spectrum of factors, including reliability, power efficiency, physical size, ageing characteristics, and yield optimization. Notably, the costs associated with technological advancements follow an exponential growth pattern, with the average expense of chip design reaching approximately 300 million dollars at the 7-nanometer scale—six times more than that at the 28-nanometer scale. It also highlights the financial investments required to develop chips at increasingly advanced technological nodes.

What are our solutions Today?

On one hand, large and experienced engineering teams are required and distributed to the various continents worldwide to meet the growing need for high quality design. For example: there were 116,000 technical employees working at Intel by 2021. The US government is expecting a growth of 89,000 US-based design workers by 2030. This will contribute significantly to the advancements in semiconductor technology. On the other hand, the utilization of Electronic Design Automation (EDA) tools plays a key role in streamlining the

¹ Gordon Earle Moore 3 January 1929 – 24 March 2023 was an American businessman, engineer, and the co-founder and emeritus chairman of Intel Corporation)

design process, reducing complexity, and saving time.

Furthermore, the industry has witnessed a notable emphasis on the reuse of validated and packaged functions. This trend underscores the growing efficiency and optimization strategies within the semiconductor design landscape [8].

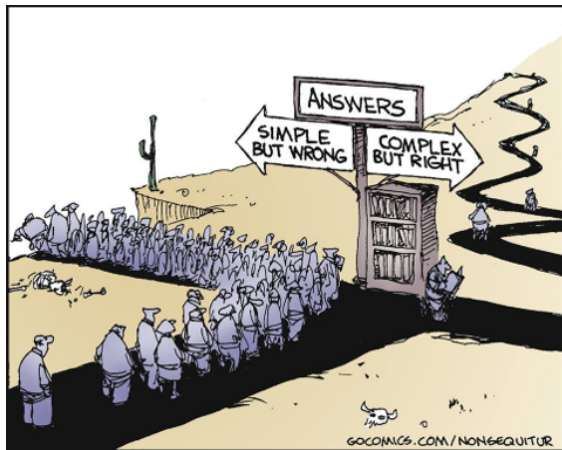


Figure 1 How to deal with complexity?

How EDA tools deal with complexity?

The evolution of tools faces considerable challenges in keeping pace with the escalating complexity inherent in semiconductor design, thereby contributing to a notable surge in design costs. The limitations of Electronic Design Automation (EDA) tools become more pronounced as design complexity intensifies, introducing a level of unpredictability to the quality of the design.

The solutions generated by these tools often need to be better and near optimal. The associated margins remain unknown, fostering a landscape where trade-offs between design technology, quality, cost, and time become increasingly intricate. Particularly with large chips, the influence of variability and technological phenomena adds layers of complexity, rendering designs more reliant on the experiences and insights of individual designers. The feedback loop on design choices, critical for achieving functional silicon, requires frequent manual interventions in the design flow, further complicating the process.

The tools themselves, though powerful, are complex to use and to tune, requiring the employment of numerous strategies to navigate the intricacies of design complexity.

Consequently, the design time becomes a variable that is difficult to predict, and convergence with hard constraints becomes a challenging task. The sheer volume of data generated in the design process adds an additional layer of complexity, presenting difficulties in storing and parsing the data and selecting the most pertinent information. As a result, larger investments are required, not only in the tools themselves but also in hardware emulators and FPGA prototypes, underscoring the multifaceted challenges posed by the evolving landscape of semiconductor design.

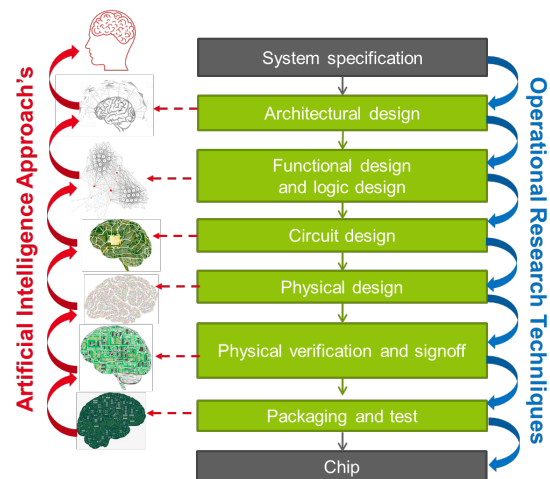


Figure 2 Electronic Design Automation Flow.

In that scope, AI techniques for Electronic Design Automation (EDA) enable facing those challenges on various aspects such as runtime and computing resources. To alleviate these challenges, Machine Learning methods are incorporated into the design process of EDA tools leveraging traditional strategies [9] as explained in next sections.

Revisiting the design Flow

The Very Large-Scale Integration (VLSI) design flow is made up of a series of steps, each of which constitutes a problem in its own right. Most of the conventional combinatorial optimization challenges emerge at various phases in VLSI design. In that field, methods based on Machine Learning techniques have demonstrated some improvement compared with traditional combinatorial optimization methods. The Design flow of digital systems can be roughly divided in six steps as depicted in ded in six steps as depicted in Figure 3 Electronic Design Automation Flow.

- 1. Architectural design:** this first step involves transforming the Integrated Circuit's High-Level description to Register Transfer Level One Description (RTL). It includes System-level Design Space Exploration (DSE) and High-Level Synthesis (HLS). The main outcomes are IC abstract specifications. The two steps are complementary.

System-level DSE determines the overall architectural parameters, while HLS performs the first premises of logic design. In these steps, several design parameters need to be fixed and evaluated. Various studies utilize simulation tools as well as predictive models for power, area, and execution time estimation. Many combinatorial optimization techniques have been used as local search, genetic algorithms, or response surface for DSE. Recently, Machine Learning strategies are used to build surrogate models and Bayesian optimization for acceleration. Collectively, these approaches contribute to advancing the efficiency and effectiveness of the IC design process, leveraging heuristics, machine learning, and surrogate models at various stages.

- 2. Functional design and logic design:** this part transforms the RTL description of a circuit to a gate-level representation in the target technology and ensure its functionality. In this step a number of transformations are applied to the design for logic optimization and minimization. The outcome is a design, represented as a netlist that is, typically, visualized as a graph of components and connections.

Mapping techniques are employed to optimize the netlists as well as Graph Neural Networks (GNN), Machine Learning and Reinforcement Learning (RL) for logic optimization on And-Inverter graphs, minimizing area under a strict timing constraint.

- 3. Circuit design:** is the phase in the design process that transitions the initial graph-based representation—comprising components and connections—from logic synthesis into a geometrical representation characterized by the shapes of materials. Once again, the use of Graph Neural Networks (Graph-NNs) is well-suited for interpreting and processing graph-based representations. This geometrical representation is often visualized as images. Leveraging the advancements in computer vision,

particularly in image classification and transformation, becomes integral to the physical design process.

- 4. Physical design:** The physical design (usually called floorplan step) phase plays a crucial role in the transformation of a design, transitioning from a graph-based representation formed during logic synthesis—comprising components and connections—to a geometrical representation consisting of material shapes called layout.

Despite five decades of research, chip floorplanning has defied automation, requiring months of intense effort by physical design engineers to produce manufactured layouts. This conversion primarily involves two key steps: Placement and Routing.

Placement, a well-established challenge in the VLSI CAD literature, has garnered increased attention due to its direct impact on physical implementation and the escalating complexity of real-world instances. Following synthesis, the placement step seeks to allocate netlist components to specific locations on the chip layout. Improved placement not only signifies enhanced chip area utilization but also holds the promise of better timing performance and routability. Despite notable research accomplishments over several decades, researchers remain dissatisfied with the efficiency of prior chip placement algorithms. The contemporary designs' colossal scale adds to the complexity of the placement process, rendering it intricate, laborious, and time-consuming.

In addition, placement has to be done considering the later routing steps, as routability is mainly influenced by the placement.

Indeed, routing is also one of the essential steps in VLSI physical design flow based on the placement assignment. Routing assigns the wires to connect the components on the chip. At the same time, routing needs to satisfy timing performance requirements and total wire length without violating the design rules. Simulated annealing heuristics among various quadratic formulations, minimum cost flow, and dynamic programming algorithms [4] have provided good solutions for different complex designs but require good expertise and iterative process over months.

Recently, AI-based techniques (mainly reinforcement and machine learning) have made big progress on these steps. As presented in [1] the reinforcement learning method automatically generates chip floorplans that are superior or comparable to those produced by humans (main metrics are power consumption, performance, and chip area) within a few hours.

Furthermore, some milestone studies have been proposed to maximize the use of GPU resources to accelerate global placement. Authors in [6] implemented the DREAMPlace placer to simulate the optimization of global placement as the Neural Network training problem and accelerates its execution on specialized HW.

5. Fabrication: lithography and mask synthesis:

In contemporary VLSI manufacturing, lithography holds a pivotal role, influencing both the printing resolution and the overall robustness of the manufacturing process. This step turns the designed circuit and layout into real objects. It involves two essential stages: mask synthesis and lithography simulation. Mask synthesis takes a layout design as input and generates a mask design with enhanced printability. Subsequently, lithography simulation utilizes the mask design to compute the printed pattern using lithography models. Since mask designs can be inherently portrayed as visual images, Machine Learning techniques, particularly Convolutional Neural Networks (CNN), are well-suited for addressing lithography challenges such as mask synthesis, modeling, and lithography hotspot detection. Furthermore, it could explore the application of machine learning in diverse manufacturing tasks, including yield estimation. Various speeds can be reached as presented in [4]. Surrogate model to estimate the yield for given design parameters is also common.

6. Physical verification and test:

Verification and testing of a circuit are complicated and expensive due to the high complexity of coverage requirements. Verification is conducted in each stage of the EDA flow to ensure that the designed chip has the correct function. However, testing is necessary for a fabricated chip. Therefore, verification and testing share common ideas and strategies while facing similar challenges from different perspectives. For instance, with the diversity of applications and

the complexity of the design, traditional formal/specification verification, and testing may only meet some demands. Mostly, verification is performed using simulations. The design is exercised with input stimuli, and its outputs are compared to golden outcomes. High coverage is reached, i.e., the fraction of functions exercised in the test. High coverage can only be achieved by many simulations with various stimuli. Two challenges arise from this. First, the required simulation time is high, and second, creating stimuli to achieve a high coverage is difficult. ML has been employed for both of these challenges.

For test set design, random generation algorithms and Automated Test Pattern Generation (ATPG) are usually used in the verification and testing stages. Their designs always need to be revised to the optimal solution. Therefore, optimizing the verification process by reducing the redundancy of the test set is intuitive.

The high complexity of chip testing/verification is another problem. More and more ML algorithms are applied in the verification and testing process to meet the coverage requirements and reduce complexity. For example, unsupervised learning may detect additional test points that improve an incomplete test plan. Some approaches also leverage AI-powered optimization software for ATPG tuning and Automatic ATPG parameter search space convergence and optimization.

Therefore, AI techniques integration is helpful at every stage of the design process, as represented in the left of Figure 3, ensuring that each step possesses its unique intelligence, along with a global intelligence facilitating communication between them to prevent falling into local minima that usually happens with heuristics approach (represented in the right side of Figure 3 in blue).

Key insight

Artificial Intelligence could play a pivotal role in reducing complex EDA flow by cleverly narrowing down the exploration space. One of its notable advantages is that it does not require a high level of tool expertise, making it accessible and more user-friendly. This democratization of expertise ensures that individuals without specialized

knowledge can benefit from AI-driven solutions. Moreover, AI could contribute to efficiency by reducing turnaround time through effective iterations and parallel runs. This acceleration in the workflow not only enhances productivity but also allows for more thorough exploration and optimization, ultimately leading to more informed decision-making.

New trends for optimization methods

Hybridization between combinatorial optimization and machine learning

The community of combinatorial optimization researchers is interested in developing exact or approximate methods to solve discrete problems that can be formulated with integer or binary variables. A cost function must then be optimized in a fixed space described using a set of equations.

The problems considered are usually NP-hard; thus, they are in theory impossible to solve exactly and efficiently. This is why approximate algorithms are usually considered to handle large instances, it is necessary to consider heuristics which build approximate solutions.

Metaheuristics were developed from the 1960s and perfected to concretely and easily solve large classes of combinatorial optimization problems: among them, genetic algorithms, simulated annealing, ant colonies, taboo methods, etc.

In 2022, researchers in combinatorial optimization and machine learning met in Dagstuhl [3] to act on the emergence of data-driven combinatorial optimization: the goal was to proceed with a hybridization between machine learning methods and those of combinatorial optimization to improve both approaches. The idea is to develop a mixed approach that will achieve the best of both worlds.

Researchers in combinatorial optimization have identified several problems when transposing their methods to the field of machine learning: among them, they identified the scalability of machine learning methods. Indeed, the optimization methods generally do not support large data sets as in machine learning. Another problem is that the evaluation of objective functions using ma-

chine learning methods does not guarantee a distance to the optimum of the solutions obtained; moreover, the results obtained are not explainable.

On the other hand, a set of techniques from machine learning should in the future benefit the solution of combinatorial optimization problems and integrated directly into Mixed Integer Linear Programming (MILP) solvers.

For example, machine learning oriented processing should make it possible to improve the choice of variables to consider in the exploration phases for the resolution of a problem expressed using a MILP solver. Likewise, data preprocessing should also make it possible to fix the values of a set of secondary parameters and thus reduce the size of the systems to be solved [4].

Using generative AI for the design flow

The tools developed by artificial intelligence researchers make it possible to completely free us from models. The existence of LLM (Large Language Models) conversational robots such as ChatGPT, available free of charge for a large audience, has a huge impact in the academic field, health, employment, etc.

The impact of this technology in the field of optimization of embedded systems is to be studied. LLMs work best when used interactively. Thus, MyCrunchGPT [5] is an overlay to ChatGPT that provides an interactive natural language-based environment to guide the user through the process of designing an optimized solution.

The authors show how this class of tools can solve various engineering problems. We find in the literature some examples of circuit creation based on an LLM: for example, Blocklove et al. [2] have created a new microprocessor architecture based on an 8-bit accumulator based and on real hardware constraints.

What Changed?

Some studies leverage AI to automate design tasks that rely heavily on human expertise and efforts. Such evolution represents a workforce development that it needed to support the semiconductor design and EDA industries. In 2023, there was a remarkable progress on ML. Moreover, the availability of much more computational power made it possible to learn ("train") on much more data

and to use bigger and "deeper" models that are more capable to solve problem.

Main Obstacles

The main obstacle identified is the lack of collaboration between the researchers and the engineers from industry and from academia. Sharing methods and results is often impossible because the software used is often seen as black boxes to protect intellectual property; the data also remains the property of the companies, which refuse their distribution, even restricted, for competitiveness reasons.

Interoperability of models and tools

Sharing CAD tools codes is currently impossible. Existing design tools come from various sources: some are open-source, and others are black boxes. The partners come from the industry or are academic researchers in optimization or ML—the tools they use come from different scientific cultures, complicating their communications. More than simply interfacing these tools with libraries dedicated to Machine Learning is also often impossible. In addition, the models used, and the results obtained are rarely disseminated by the industry for intellectual property reasons.

Availability of Training Data

One of the current difficulties is the generation of training data. Indeed, deep neural networks require significant data for their training. Furthermore, this data must come from various designers to obtain good solutions in inference.

The technical difficulty of sharing these data is a direct consequence of the lack of interoperability of the models: the models used differ depending on the customers. Moreover, industry often are reluctant to exchange numerical values for confidentiality reasons. The lack of data constitutes then an obstacle for academic research, which needs access to real-life industrial data, and for the development of new methods in the industrial context associated with the lack of knowledge of optimization methods and ML tools.

Conclusion

Artificial Intelligence could be pivotal in reducing complex EDA flow by intelligently narrowing the exploration space in addition to combinatorial optimization approaches. One of its notable advantages is that it does not demand a high level of user tool expertise, making it accessible and user-friendly. This democratization of expertise ensures that individuals without specialized knowledge can benefit from AI-driven solutions. Moreover, AI could improve efficiency by reducing turnaround time through effective iterations and parallel runs. This acceleration in the workflow enhances productivity and allows for more thorough exploration and optimization, ultimately leading to more informed decision-making.

As presented here, general algorithms from the combinatorial optimization field plus a growing collection of new models from machine learning and AI optimization methods, as well as more computational power, made it possible to learn (train) on much data and use bigger and deeper models contributes on one side to reduce design cycle time to be smaller and more automatic and on the other side being able to build more customized chip easily. Thus, it represents an important potential for hardware design.

However, many challenges arise. One of them is the Environmental footprint of AI, which is not negligible. For example, training GPT-3 (which has 175 billion parameters, consumed 1 287 megawatt hours of electricity and generated 552 tons of carbon dioxide [7]).

Open questions and ongoing challenges

These Decision-aid strategies rely on three essential ingredients:

1. streamlined algorithms, to provide optimal or good solution.
2. the accessibility of data (particularly pertinent for AI-driven methodologies). Note that availability of data is a major step forward for the research community and will be gratefully acknowledged.
3. the lack of interoperability between software of EDA and those used in ML, as well as their opacity. This obstacle could be removed by making available open-source software whose performance would be comparable to those used in the industrial environment.

4. hardware architecture coupled with accelerators, which need to be fast enough to train data in reasonable time. Thus, facilitating the timely execution of models.

References

- [1] Chowdary, D., & Sudhakar, M. S. (2023). Multi-objective Floorplanning optimization engaging dynamic programming for system on chip. *Microelectronics Journal*, 140, 105942.
- [2] Blocklove, J., Garg, S., Karri, R., & Pearce, H. (2023). Chip-Chat: Challenges and Opportunities in Conversational Hardware Design. *arXiv preprint arXiv:2305.13243*.
- [3] Frejinger, E., Lodi, A., Lombardi, M., & Yorke-Smith, N. (2023). Data-Driven Combinatorial Optimisation (Dagstuhl Seminar 22431). In *Dagstuhl Reports (Vol. 12, No. 10)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.
- [4] Kalla, D., & Smith, N. (2023). Study and Analysis of Chat GPT and its Impact on Different Fields of Study. *International Journal of Innovative Science and Research Technology*, 8(3).
- [5] Varun Kumar, Leonard Gleyzer, Adar Kahana, Khemraj Shukla, and George Em Karniadakis. MyCrunchGPT: a LLM assisted framework for scientific machine learning. *Journal of Machine Learning for Modelling and Computing*, 4(4):41–72, 2023
- [6] Yibo Lin, Shounak Dhar, Wuxi Li, Haoxing Ren, Brucec Khailany, and David Z Pan. Dreamplace: Deep learning toolkit-enabled gpu acceleration for modern VLSI placement. In *Proceedings of the 56th Annual Design Automation Conference 2019*, pages 1–6, 2019.
- [7] Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., & Dean, J. (2021). Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.
- [8] L. Zaourar & N. Ventroux, Machine Learning for Design Space Exploration of CPS”, Workshop Machine Learning for CAD, DATE 2019, Florence, Italy.
- [9] Rapp, M., Amrouch, H., Lin, Y., Yu, B., Pan, D. Z., Wolf, M., & Henkel, J. (2021). MLCAD: A survey of research in machine learning for CAD keynote paper. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 41(10), 3162-3181.

Alix Munier Kordon is a full professor of Computer Science of at Sorbonne Université and member of LIP6, Paris, France.

Lilia Zaourar is an expert in the research and technology department at CEA (the French Atomic Energy Commission).



The AI scene in Europe is dynamic and evolving.

State of artificial intelligence in Europe

by Jack Vernon

It is important to position Europe in the race of artificial intelligence. This article gives some hints on this topic and is based on surveys and analysis done essentially in 2023. It is by no means an exhaustive view of the position of Europe in the field of artificial intelligence but does offer significant detail and insight into each of the topics. The data used in this report comes from three IDC surveys [1],[2],[3], and two IDC forecasts [4],[5].

Market Developments and Dynamics.

European AI Adoption

The adoption of AI technologies is already making good progress across Europe, with many organisations reporting the introduction of various forms of AI. IDC considers AI to fall into four categories, which are as follows:

- **Intelligent Process Automation:** is a class of software designed to automate or augment manual repetitive tasks.
- **Predictive AI:** Analysis of large data sets to identify long term patterns in behaviour and detect changes (e.g., digital twins and threat detection).
- **Descriptive AI:** Analysis of images or event data streams so people and things can detect, analyze, and act (e.g., machine vision).
- **Generative AI:** Creates new content/code using previously created content/code (e.g., ChatGPT and developer co-pilots).

As seen in Figure 1, Generative AI leads in terms of the number of organisations saying they already use the technology, followed closely by IPA. Descriptive and Predictive AI have similar volumes of respondents stating they are already using them. Given the ease of use and low cost of general-purpose AI-driven conversational systems like ChatGPT and Bard, it is perhaps not surprising to see Generative AI leading in terms of the AI technology most organisations already use. Staff within an organisation can use generative AI assistants free of charge, and so in many cases, businesses might use generative AI informally, yet significantly.

In contrast, deploying forms of Descriptive or Predictive AI requires a degree of configuration specific to a business use case. This requires expertise and, unlike a generative AI assistant, can involve using technology that isn't free. So, although other forms of AI have been available to businesses for a longer period, respondents to the survey report lower levels of adoption.

The drive to expand the use of AI technologies is strong across all categories. Descriptive and predictive AI are set to grow faster than both IPA and generative AI, although they are helped by starting from a lower base of existing adoption. In all categories, respondents planning to introduce AI

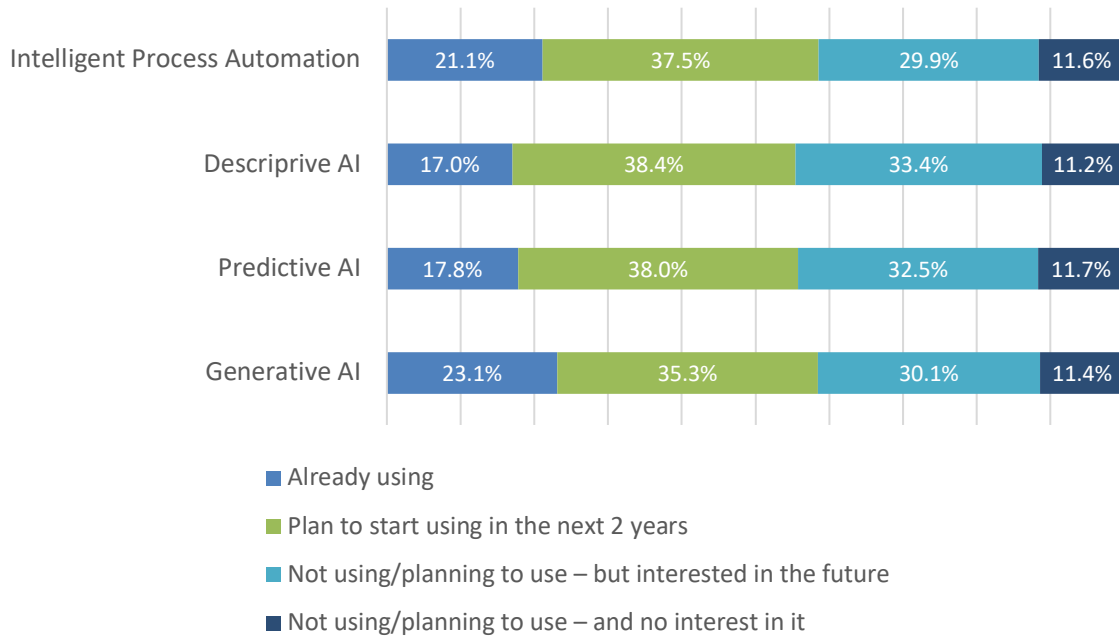


Figure 1: Collected answers to the question "Is your Organisation Currently Using or Planning to Use a Form of AI Technology?". From [2]

technologies in the next 2 years are at least 150% larger than the group already using the AI technology.

In terms of adoption by countries and subregions, several key differences exist between countries. Nordic respondents lead in current Generative AI

usage but have more limited plans for the future introduction of Intelligent Process Automation. Spain showcases consistency in its interest across all AI technologies, with high current usage and robust future adoption plans. Poland, despite lower current usage rates in several categories, shows a strong ambition for the future, notably in

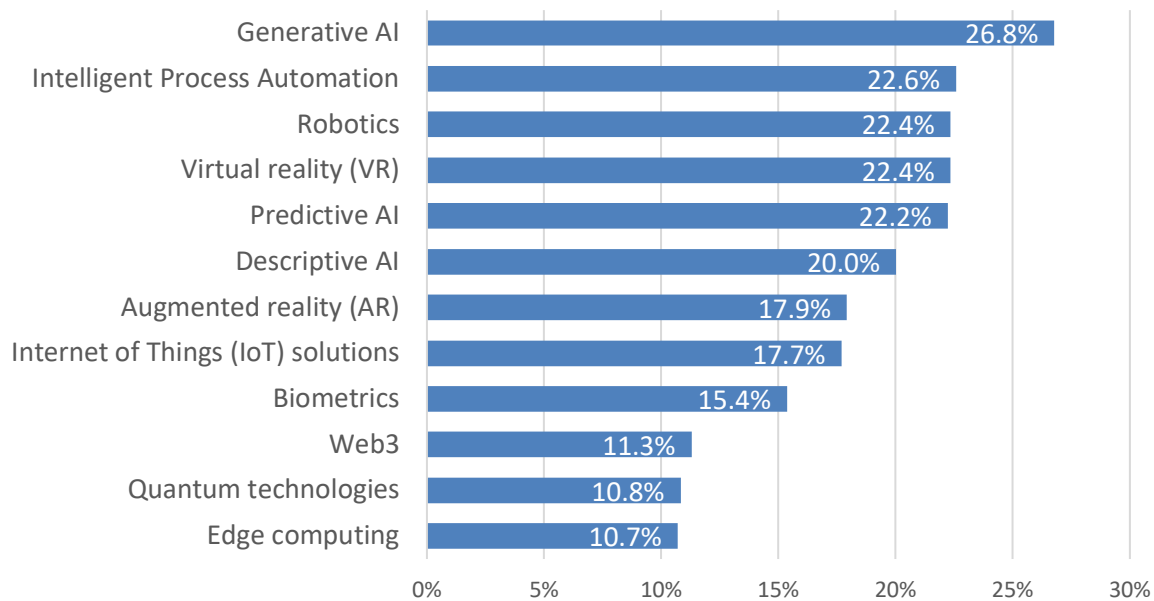


Figure 2: Collected answers to the question "Which of these technologies do you believe are most likely to have a significant transformative impact on your organization? [Choose up to 3] Top 12 Responses". From [2]

Intelligent Process Automation and Generative AI. Czechia stands out with high current usage of Predictive AI but appears more reserved in its future adoption plans compared to other nations. Germany and France feature moderate adoption without drastic highs or lows. However, France's current usage in Intelligent Process Automation is notably higher. The UK is set to increase its focus on Predictive and Descriptive AI in the coming years. Germany, while having lower current usage in Intelligent Process Automation, has plans suggesting a rise in the upcoming years.

When asked about their perception concerning the transformative impact of emerging technologies, respondents rated Generative AI as the most impactful by a significant margin, followed closely by other AI-related technologies, see Figure 2. The combined potential impact of AI technologies overshadows many other categories, showcasing the growing emphasis on data-driven decision-making in the modern business landscape. Technologies like Blockchain, Quantum technologies, and Web3 received relatively less attention, with scores around the 9-11% range, possibly due to their niche applications or a lesser understanding of their potential impact. Virtual Reality (VR) and Robotics are parallel in significance to AI and IPA. Descriptive AI ranks lowest of the AI categories, potentially due to its technological limitations and the challenges in deploying it in live

environments without substantial computational resources.

Influencing Factors

Considering the factors most likely to influence the adoption of emerging technologies within organisations, respondents highlighted a combination of driving and inhibiting factors. Top inhibitors included cybersecurity, economic stability, and the skills shortage. Key drivers included digital innovation, sustainability requirements, and the development of new digital business models. Among these concerns, cybersecurity alone garnered over 30% agreement from respondents. Six influencing factors received agreement from between 30–20% of respondents, and a further six factors received agreement from between 20–15% of respondents. The lower tier of responses tended to be relatively condensed, with no single factor standing out in terms of relatively poor performance. The question addresses emerging technology generally, so the specific influencing factors may vary slightly for AI.

AI Spending

The latest Worldwide Artificial Intelligence Spending Guide (V2 2023) published by IDC shows that artificial intelligence (AI) spending in



Figure 3: Collected answers to the question "What Factors Most Influence your Organisations Decision to Implement an Emerging Technology". From [2]

Europe will reach \$34.2 billion in 2023, representing 20.6% of the worldwide AI market (see Figure 4). AI spending in Europe will post a 29.6% compound annual growth rate (CAGR) between 2022 and 2027, slightly higher than the worldwide CAGR of 26.9% for the same period, with spending expected to exceed \$96.1 billion in 2027.

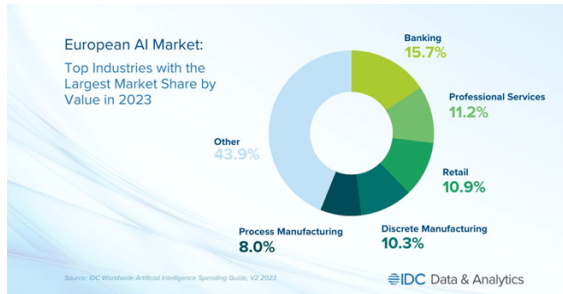


Figure 4: European AI Spending Forecast, by Industry. From IDC Worldwide Artificial Intelligence Spending Guide, V2 2023

European AI Strategy - Investment intentions

European organisations have indicated in a survey [1] about the use and adoption of Generative AI their investment expectations for different AI technologies. A smaller proportion of investment is anticipated for generative AI compared to other categories; however, this still represents a significant increase from the previous year, when this figure was significantly lower. Descriptive AI is expected to receive the largest share of investment. Predictive AI will also see significant investment, representing over a third of the anticipated AI investment allocations.

The data provides an interesting contrast with the indications of adoption gathered from the emerging technology survey in the previous section. Although respondents to the emerging tech survey suggested generative AI adoption would be similar to or greater than that of other AI technologies, spending intentions differ. Companies in the Generative Arc survey indicate they anticipate allocating significantly larger amounts to more mature AI technologies like Predictive and Descriptive AI.

Top Use Cases

The emerging technologies survey [1] also included a question covering the popularity of AI use cases in terms of adoption, and planned adoption in the next 24 months. Considering the top 10 use cases, conversational AI platforms for automated customer service are the most adopted AI technology among the listed use cases. There is a strong emphasis on automation across various domains, such as IT operations, sales processes, and knowledge work. AI use cases that have predictive capabilities, such as asset operations and threat intelligence, indicate a trend towards investing in the anticipation of challenges and opportunities.

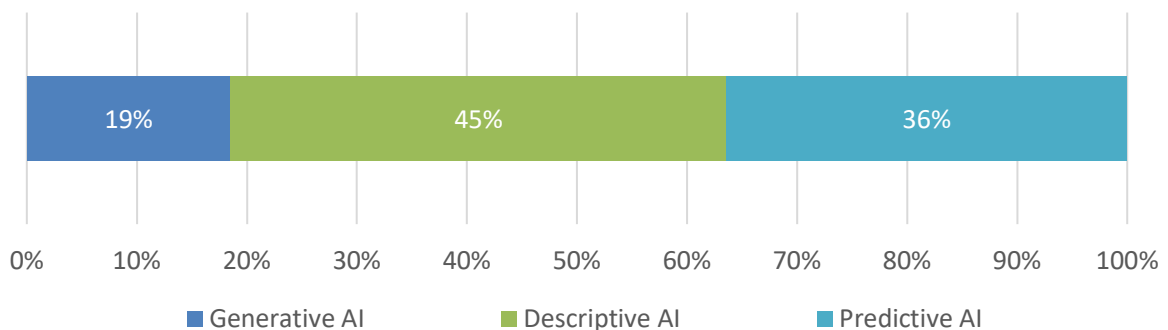


Figure 5: Collected answers to the question " Mean Summary of AI Investments - Thinking about your organization's investment in all AI-related development, data, and infrastructure assets right now, what is your best estimate of how these investments will be allocated over the next 18 months?". From [1]



Figure 6: Collected answers to the question "In which of the following areas does your organization use or plan to use Artificial intelligence in the next 24 months? [Choose all that apply] - Top 10". From [2]

Measured Benefits

The measured benefits of AI reflect the business outcomes of popular use cases. The data indicates that organisations are reaping substantial benefits from implementing Artificial Intelligence systems, with significant improvements notably in process quality and efficiency. A significant majority report enhanced time management, increased customer satisfaction, and revenue growth. Furthermore, AI's impact on cost reduction is equally significant as its role in reducing environmental footprints, underscoring its influence on operational efficiency and sustainability. The advancements have

also spurred the launch of new products and services, affirming AI's role as a driver of innovation. A minimal number of respondents report not seeing benefits yet, indicating a broadly positive impact across various sectors.

Mature Enterprises

The gap between organisations regarding their strategic maturity on AI is also distinguished by factors concerning how their AI operations are organized, and not exclusively reporting of its introduction. A crucial indicator of maturity is how AI work is distributed within a business. If most AI

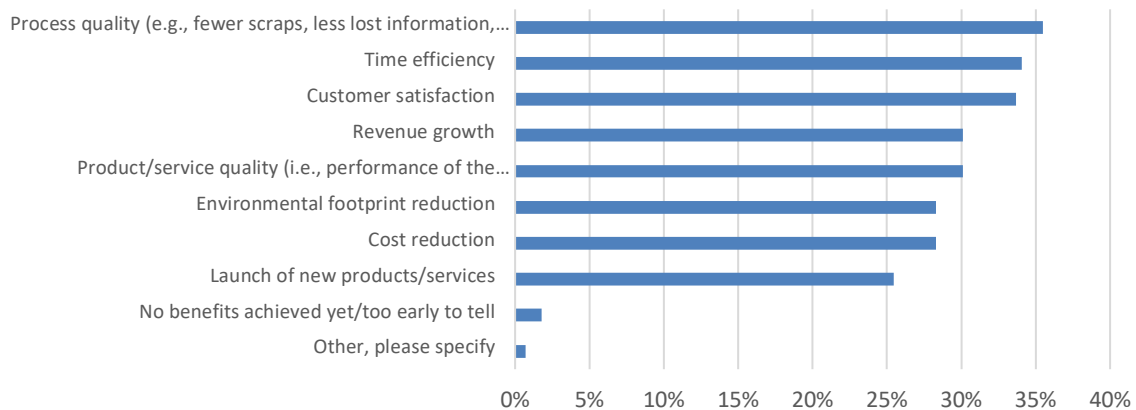


Figure 7: Collected answers to the question "Which of the following measurable results have been achieved by your organization thanks to the implementation of Artificial Intelligence (AI) emerging technologies? [Choose all that apply]". From [2]



Figure 8: Collected answers to the question "What Processes has your organization introduced to ensure safe and ethical development and use of AI within your organization? [Choose All That Apply]". From IDC Europe, *Emerging Technologies Survey, VIDEA July 2021, Subsample: AI Use or Plan to Use*

activity is centralised to a single team within the IT department, then typically, these businesses are less mature than those that have AI practitioners and data scientists distributed throughout distinct business teams and departments. Centralised AI teams often face resistance from business leaders and may have resource constraints. They may need to bid for work internally, which often leads only to Proof of Concept (PoC) projects instead of live deployments. Leaders of business teams should be empowered to hire data science and AI practitioners directly, allowing them to interact with key decision-makers and have proximity to business problems.

European organisations need internal policies governing the development and deployment of AI technology. AI presents business and ethical risks which are increasingly transforming from reputationally to legally damaging. Without informed internal policy on AI, executive team members responsible for managing AI policy, and a working group reviewing policy improvements, an organisation may struggle to keep pace with AI's regulatory and ethical developments. Technology systems can support companies in rationalising AI development and implementation, ensuring best practice adherence. Systems related to AI explainability, placing humans in the loop, organising data science activity, and MLOps can all support organisations in meeting AI ethics standards.

When the European Emerging Technologies survey was conducted in 2021, those respondents already using AI were asked about their strategies on AI ethics issues. Some organisations had

executive oversight in place and potentially a governance framework for developing AI. Very few respondents undertook more involved steps such as conformity assessments or ethical auditing of AI systems.

Conclusion

The landscape of AI adoption and trends in Europe reflects a dynamic and evolving environment, marked by a strong drive to integrate various forms of AI across multiple sectors. European organisations are actively embracing AI technologies, with a notable lean towards Generative AI, as evidenced by widespread use and interest. Although as we'll see in the following section, Generative AI is not as popular in Europe as other global regions. This is followed closely by Intelligent Process Automation (IPA), Predictive AI, and Descriptive AI, with each having its distinct application and growth trajectory.

Investment intentions in Europe show a cautious but strategic approach, favouring more mature AI technologies like Predictive and Descriptive AI over Generative AI. This trend likely stems from a combination of factors, including data privacy concerns and the anticipation of future regulations. Despite this, there's a clear recognition of AI's transformative impact, particularly in Generative AI, which is viewed as the most impactful emerging technology.

Differences in adoption rates and focus areas are evident across various European countries and subregions. The measured benefits of AI in

Europe include enhanced process quality and efficiency, improved time management, increased customer satisfaction, and revenue growth. Notably, AI's role in cost reduction is as significant as its contributions to reducing environmental footprints, highlighting its dual impact on operational efficiency and sustainability.

Strategic maturity in AI varies across European organisations, with more mature entities distributing AI roles across different business teams, as opposed to centralizing them within IT departments. This maturity is not just about the adoption of technology but also involves internal policies and ethical considerations. European organisations are increasingly focusing on AI governance, ethical standards, and compliance with regulatory developments.

Overall, Europe's approach to AI adoption and its trends demonstrate a balanced blend of enthusiasm and caution, with a strong emphasis on ethical considerations, regulatory compliance, and strategic investment in various AI technologies.

The European position in AI, a comparison between regions

The following section will compare the position of the European AI market with that of other global regions. European businesses are following similar trends to those displayed in other regions. However Europe is more cautious in terms of AI investments and approach to generative AI.

Revenue Comparison

As shown by Figure 9, according to IDC's AI Tracker, EMEA is predicted to be the fastest-growing global region in terms of CAGR during the 2022–2027 period, slightly surpassing the APJ and Americas subregions. In terms of EMEA's share of the overall AI market, it will lag behind the Americas but is expected to gain some ground on the region throughout the period. Europe, as the largest revenue contributor in the EMEA region, exhibits growth and market share trends similar to those seen across Europe.

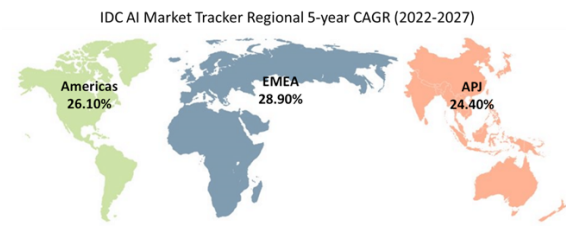


Figure 9: Regional Compound Annual Revenue Growth (CARG) Comparison. From IDC's Worldwide Semiannual Artificial Intelligence Infrastructure Tracker 2023

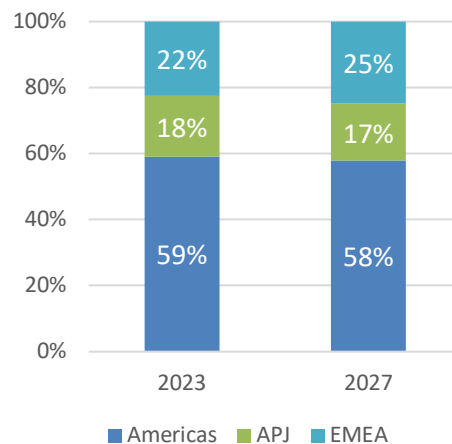


Figure 10: Share of AI Revenue by Global Region. From IDC's Worldwide Semi-annual Artificial Intelligence Infrastructure Tracker 2023

Investment Intentions

When asked about their investment intentions for AI in the Generative AI Arc survey, European respondents expected to invest more in predictive AI than their counterparts in North America and Asia Pacific, but less in generative AI. Europe's investment in Descriptive AI is comparable to that of North America but lower than that of META (Middle East, Turkey, and Africa). European respondents' more restrained investment expectations in generative AI indicate a more cautious approach than other regions to the category, possibly influenced by data privacy concerns and the anticipation of future regulation.

Business Objectives

Figure 12 comes from a Global 2022 survey of AI users; respondents were asked which business objectives are the highest priority for AI projects.

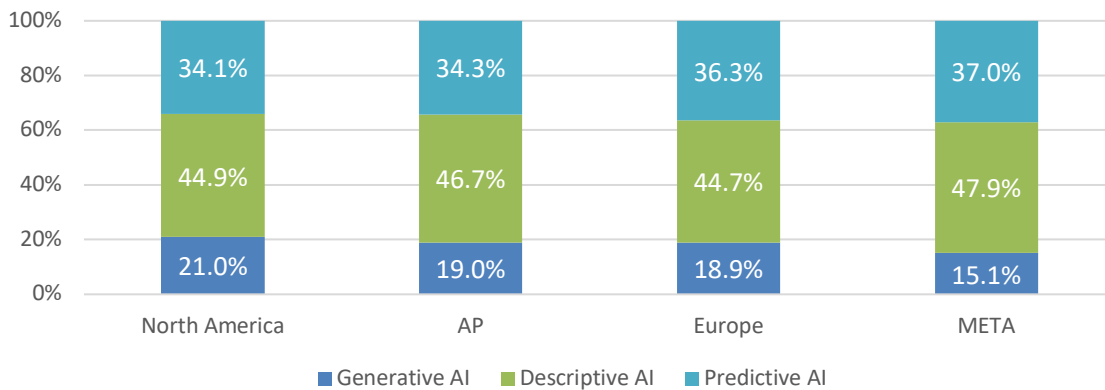


Figure 11: Collected answers to the question "Thinking about your organization's investment in all AI-related development, data, and infrastructure assets right now, what is your best estimate of how these investments will be allocated over the next 18 months?". From [1]

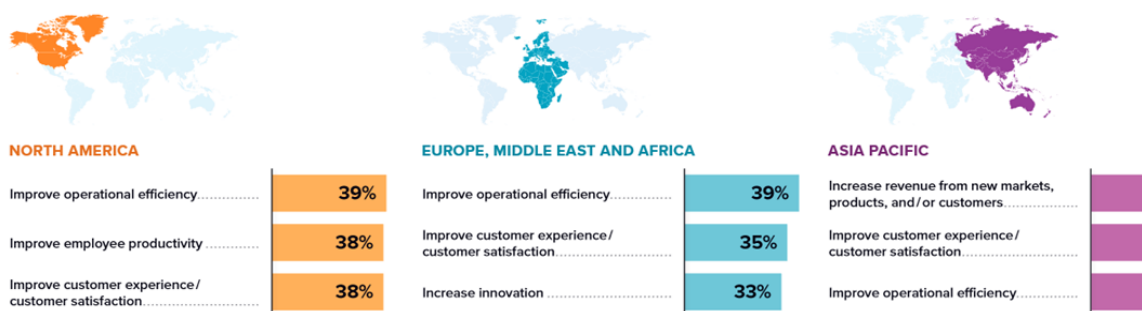


Figure 12: Collected answers to the question "What are your primary business objectives for using AI for your projects/initiatives?". From IDC's AI Strategies View 2022, May 2022

The objectives that resonated most with EMEA respondents were aligned with those of respondents from the Asia Pacific (AP). EMEA respondents were less enthusiastic about objectives beyond improving operational efficiency than those in other regions. In contrast with EMEA and AP, North American respondents ranked improving employee productivity as the second top priority, a factor EMEA businesses may be more cautious about openly communicating or prioritizing in AI projects.

AI project Failures

Considering companies that experienced AI failures, EMEA respondents indicated that they experienced higher rates of project failure than other regions. Globally, companies across all regions cited 'AI technology not performing as expected or as promised' as the primary reason for project failure. EMEA respondents also identified 'lack of

skilled personnel and staff' as the second main reason for project failure, whereas this reason did not rank in the top three in other regions. The survey results suggest that EMEA may be facing a more acute skills shortage than other global regions, potentially leading to higher rates of AI project failure.

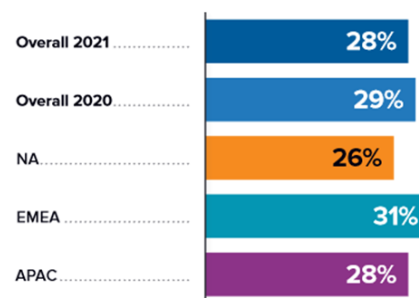


Figure 13: Percentage of Respondents that noted AI Failure. From IDC's AI Strategies View 2022, May 2022

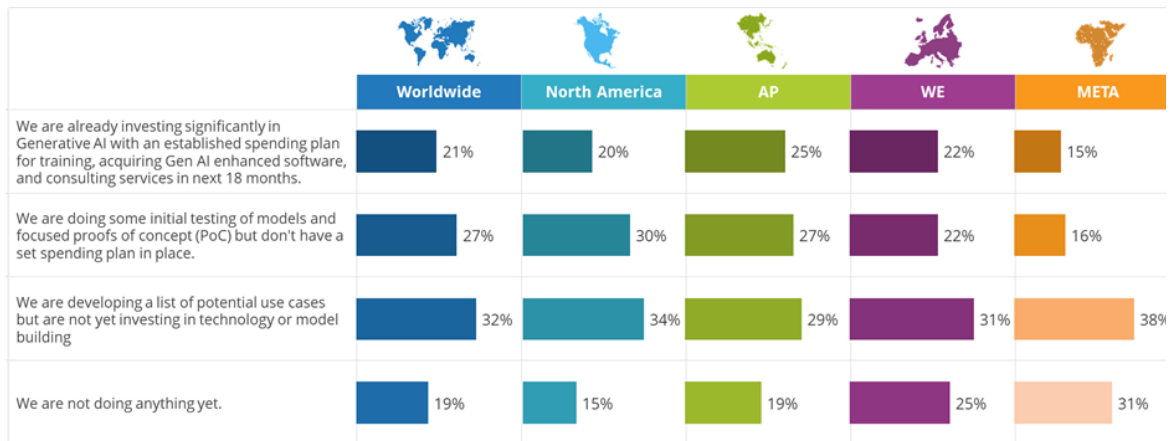


Figure 14: Collected answers to the question "What is your organization's current state of evaluating or using Generative ". From [1]

Ecosystem

Europe's position on AI is similar to its stance on other digital technologies. It consistently provides industry talent, thanks to numerous top academic institutions, and benefits from several high-income countries, often powered by mature financial services and industrial sectors. However, Europe has had limited success in retaining local ownership of its most promising AI companies. Google's acquisition of DeepMind in 2014 set a precedent, and since then, Europe's most promising AI software businesses often ended up under the ownership of US-based hyperscalers.

There are exceptions. The AI platform vendor Dataiku continues to grow an impressive business portfolio, with commitments from the founding team to maintain independent. Advanced AI consultancy InstaDeep also bucked the trend by signing an acquisition deal with German pharmaceutical company BioNTech.

Recent developments in the Generative AI space have created a new market category where European businesses can compete. Several European startups have launched in this space, from generative language model providers like Aleph Alpha and Mistral AI to image generation enterprises such as Stability AI. It remains to be seen whether European vendors will take a commanding lead in generative AI technologies, although specialization in certain localized languages could help them become prominent players in specific European markets.

Generative AI Preferences

Europe appears to be adopting a different approach to generative AI technology compared to other regions. This section will explore data points from [1], highlighting how European respondents compare with those from other regions. The survey suggests that Europe is neither significantly behind nor diverging greatly in its approach compared to other global regions.

When asked about their organisation's 'current state of evaluation or use of Generative AI' technology, a larger proportion of European respondents than in other regions indicated that their company was currently not engaged in AI. Europe was also 5% below the global average in terms of respondents who agreed their company was working on a generative PoC project. Although the number of respondents who agreed their company had already implemented a Generative AI investment plan was in line with the global average, the survey results indicate European respondents are less advanced in developing Generative AI investment plans than other global regions.

Considering Figure 15 and the question, "What types of Generative AI models are organisations using or testing?" 81.9% of European respondents are using private versions of models for experimentation. Companies in Europe exhibit a clear inclination to safeguard their intellectual property. As a result, Europe is more likely than other regions to overlook third-party generative AI applications and public versions of generative AI models.

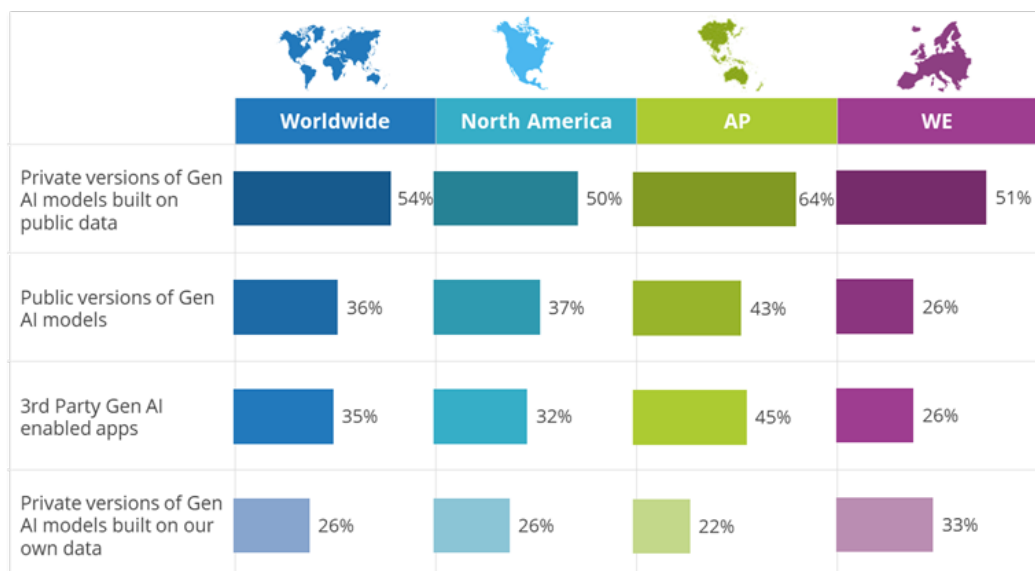


Figure 15: Collected answers to the question "What Generative AI model approach is your company using or investigating?". From [1]

European organisations show a preference for experimenting with GenAI models that are pre-trained on public datasets. Websites such as HuggingFace and PapersWithCode have provided the data science community with well-labelled public datasets that can be conveniently incorporated into a training pipeline. The cumulative total of generative AI approaches being tested by respondents is 135%, indicating that many are experimenting with multiple generative AI approaches.

Conclusion

The European AI market exhibits unique characteristics and trends that differentiate it from other global regions. While the EMEA region is projected to experience the fastest growth in terms of Compound Annual Revenue Growth (CARG) during 2022-2027, it still lags behind the Americas in overall AI market share. Investment patterns in Europe reflect a cautious approach, particularly in generative AI, likely influenced by data privacy concerns and anticipated future regulations. This contrasts with more aggressive investments in predictive and descriptive AI technologies.

In terms of business objectives, European organisations prioritise operational efficiency but show less enthusiasm for goals like improving employee productivity, a priority in North America. This reflects a more reserved approach to AI's role in workforce management. Additionally, the EMEA region faces a pronounced skills shortage, which

has led to higher rates of AI project failures compared to other regions, indicating a critical area for improvement.

The European AI ecosystem, while rich in talent and resources, struggles to retain local ownership of its most promising AI companies, often losing them to US-based hyperscalers. However, there are notable exceptions, and recent developments in generative AI have opened new opportunities for European businesses. European organisations are more inclined to use private AI models for experimentation, safeguarding intellectual property, and showing a preference for GenAI models pre-trained on public datasets. This indicates a trend towards a more controlled and proprietary approach to AI development in Europe.

In summary, while Europe aligns with global trends in some respects, its cautious investment strategy, focus on operational efficiency, challenges in talent retention, and distinct approach to generative AI set it apart from other regions. These factors underscore Europe's unique position in the global AI landscape.

AI Regulation in the EU and other world regions

In recent years, the field of AI has grown fast, shaping industries and affecting our ways of living and working. IDC expects the European AI market to reach \$72 billion by 2026 (IDC Worldwide Artificial Intelligence Spending Guide - Forecast 2023). As AI technologies and systems become

progressively more integrated into people's lives, concerns regarding their ethical implications, potential risks, and misuse have started emerging.

Key Issues Around AI

- AI may provide lower-cost solutions and enhance productivity, but due to its “black box” nature, it has garnered much attention for its ethical implications.
- Privacy and Consent: Concerns over whether appropriate consent has been obtained in training datasets.
- Biases and Toxicity: AI may produce biases such as racial or gender-based discrimination due to incomplete datasets or reproduction of human biases within the data.
- Harmful Content: AI may provide explicit content, propaganda, or misinformation.
- Security: AI may not be safeguarded against manipulation by third parties, may be susceptible to “Shadow AI,” or may be more vulnerable to data breaches.
- Accountability: AI may not have a “human in the loop” to sufficiently monitor, test, and update the system.
- Lack of Explainability: AI model may be a “black box,” meaning even its developers don’t understand how certain outputs were produced, causing security concerns.

The Broader AI Risk Landscape

On a larger scale, these questions around ethics, accountability, and transparency will have vast implications on society.

- Amplification of bias or discrimination
- Misinformation and disinformation
- Erosion of privacy
- Misuse for political or geopolitical aims

IDC predicts that by 2028, taking cue from the EU’s AI policy, 60% of worldwide national governments will adopt a risk management approach in framing their AI and GenAI policies. (IDC FutureScape 2024: Worldwide National Government)

European Union: Overview of AI Regulatory Landscape – The EU AI Act

The European Union embarked on an ambitious journey to regulate the deployment and development of AI systems through the introduction of the EU AI Act in 2019, highlighting the need for a thorough regulatory framework.

In April 2021, the European Commission submitted a detailed proposal of its plan. After the proposal, the Commission adopted a “general approach” on a set of harmonized rules on artificial intelligence. The latest developments of the technology, particularly generative AI, caused some delays in the final discussion of the legislation as new amendments were deemed necessary, but on May 11, 2023, the European Commission committees approved the proposed amendments in what was considered the first milestone vote on the EU AI Act. The plenary vote on June 14, 2023, approved with a large majority by the European Parliament, signals the beginning of the final phase of the legislative process — the “trialogue.” Here, high-level negotiations between the EU Parliament, Council, and Commission are expected to last until the end of 2023¹. If negotiations are successful, the EU AI Act will come into effect in June 2024, with a two-year transition period for AI developers and providers to adjust. Ultimately, it will be Member States who will enforce the regulation.

The AI Act

A comprehensive regulatory framework designed to maintain a balance between innovation and ethical use of AI technologies. It aims to safeguard EU fundamental rights and values while fostering advancements in AI research and application.

The legislation emphasizes the improvement of data quality, promoting transparency in AI operations, and enforcing human oversight to ensure accountability and ethical use. It also addresses liability concerns, particularly in critical sectors such as healthcare, finance, education, and energy, emphasizing the need for responsible AI integration in these domains.

Scope and targets: The AI Act targets providers placing AI systems in the EU markets, irrespective of whether those providers are based in the region or in a third country; users of AI systems physically established within the Union; providers and

¹ This article was written mid November 2023.

users of AI systems based in a third country, but whose systems produced outputs based in the Union; importers and distributors of AI systems; manufacturers placing AI systems on the market with their product and under their trademark; authorized representatives of providers based in the EU.

Risk-based approach: The regulation identifies four risk categories for AI applications, and applies different restrictions and obligations on system providers and users, depending on the category of the application in question:

- **Unacceptable risk.** This category targets applications that involve subliminal practices, exploitative activity, or social scoring systems by public authorities. It covers cognitive behavioural manipulation of people or specific groups; people classification based on behaviour, socioeconomic status or personal features; and real-time and remote biometric identification, such as facial recognition. Such applications will be banned. This risk category allows for some exceptions, creating a sort of grey area. For example, "post" remote biometric identification is permitted via court approval if it occurs with a significant delay to prosecute serious crimes.
- **High risk.** Applications related to education, healthcare, and employment (such as CV scanning, ranking job applicants) will be subject to specific legal requirements (e.g., to ensure transparency and safety of the systems and comply with the Commission's mandatory conformity requirements). It covers AI systems used in products falling under the EU's product safety legislations (toys, aviation, cars, medical devices, and lifts). It also identifies eight specific categories of AI systems to be registered into an EU database. Providers of "high risk" systems have obligations to establish quality management systems, keep technical documentation up to date, undergo conformity assessments (and re-assessments), conduct post-market monitoring, and collaborate with surveillance authorities.
- **Limited risk.** AI systems such as chatbots will be subject to minimal specific transparency obligations, for example, disclosing that interactions are performed by a machine, so that users can take informed decisions. This category covers AI systems that generate or

manipulate image, audio, or video content, such as deepfakes.

- **Minimal risk.** Applications that are not listed as risky nor explicitly banned are left largely unregulated (e.g., AI-enabled video games). Currently, this category covers most AI systems used in the EU.

GenAI Focus: The surge of Generative AI systems caused the addition of specific amendments to the AI Act. The focus here is on transparency requirements: Generative AI systems are required to disclose that the content was generated by AI; systems and models need to be designed to prevent them generating illegal content; summarizing copyrighted data use for training.

Trajectory of Regulation:

- The member states and Commission will establish a database of high risk "AI systems." The database will hold information on those AI systems being used or sold in European markets.
- AI 'providers' will be tasked with the submission of any high-risk AI system to a European database. Submissions will include information covering, training data, traceability, transparency, accuracy, potential risks presented by the system, details concerning the intended application, bias assessments.
- Member states will be tasked with designating national authorities/existing ministries to enforce the regulation and perform market surveillance.
- Fines for non-compliance range from €30m or 6% of global turnover, whichever figure is higher.

Technology and Business Impact:

- The AI Act will support AI innovation and adoption by reducing legal uncertainty, creating a level playing field for businesses, and improving the quality and safety of AI systems.
- Certain use cases or industries will be more impacted by the AI Act than others, depending on the risk level and the potential benefits of the AI application.
- The use of biometric identification technologies will fall into the 'unacceptable' or 'high risk' categories and therefore be imposed with stringent limitations. This will limit the use of

these technologies for some applications, such as security or fraud prevention, but will also provide greater protection for individual privacy and rights.

- However, the majority of current AI-related use cases will not fall under the high-risk categories. Consequently, any potential decrease in value for high-risk use cases would not lead to a dramatic drop in the overall market size.
- Larger companies have fewer concerns as they possess the financial resources to cover necessary audits and compliance costs. They will continue investing in building relationships with regulators, which could lead to long-term changes in the technology landscape.
- VCs and AI startups express apprehension that the new regulation will limit their ability to innovate and compete. The challenges they anticipate primarily involve technical complexities, compliance costs, and additional obligations associated with high-risk AI systems.

United States: Overview of AI Regulatory Landscape

Comprehensive US legislation governing use of AI is current pending discussion. A number of voluntary frameworks have also emerged. Agency-specific and state-specific legislations also exist.

- Executive Order on Safe, Secure, and Trustworthy AI (2023, published): Executive order announced in October 2023. The order requires developers of the most powerful AI systems to share safety test results and other critical info with the US government; requires that the National Institute of Standards and Technology (NIST) develop of standards, tools, and tests to ensure safety of AI systems; calls on Congress to enact bipartisan privacy protection legislation, as well as AI-fraud protection; and more.
- Bipartisan Framework on AI Legislation (2023, draft): Blueprint for a future AI Act proposing: (1) a licensing regime for sophisticated or high-risk AI, (2) an independent oversight body, (3) legal protections to tech firms for third party content, (4) transparency obligations, and (5) other protections (e.g. consumer, child, national safety).

- Digital Platform Commission Act (2023, draft): Draft bill proposing a new Federal body to oversee and regulate digital platforms, which are understood to include social media platforms and their associated algorithms and databases.
- Ensuring Safe, Secure and Trustworthy AI (2023, published): A list of 8 voluntary commitments that companies will make to promote safe, secure, and transparent development and use of AI. Recently, Adobe, Cohere, IBM, NVIDIA, Palantir, Salesforce, Scale AI, and Stability AI have made this voluntary commitment.
- Blueprint for AI Bill of Rights (2022, published): A voluntary blueprint identifying 5 principles for “automated system” design, use, and deployment: (1) safe and effective systems, (2) algorithmic discrimination protections, (3) data privacy, (4) notice and explanation, and (5) human alternatives, consideration, and fallback.

Current Key Characteristics of the US’s AI Regulation Regime

- Voluntary Frameworks vs. Full Regulation: Voluntary frameworks help pave the way for standards of conduct until legislation is fully passed, which is a lengthy process in the US. However, this is obviously dependent on companies upholding these frameworks, but given customer expectations around this, will be crucial and can offer competitive advantages to companies that are more transparent about their AI strategies.
- State-level Solutions vs. National Law: Legislation already exists, especially on a state level, though this report does not go in-depth on these. For example, Connecticut has its Artificial Intelligence Law that establishes a task force to research and make recommendations on AI. This has led to a patchwork landscape as certain states implement AI regulation more strictly than others. State-level regulations also tend to focus more on very specific use cases, such as HR programs (e.g., New York and Illinois laws).
- Innovation vs. Regulation: The US regulatory landscape is overall less developed than the EU, China, and other select countries. It is patchwork and drafts remain suspended in debate (and often embroiled in political strife). Yet, the US is home to some of the fastest-

paced innovation in the AI field - from Microsoft and Google to AI startups. Walking the fine line between innovation and regulation is critical to the US, especially given how geopolitically-conscious regulators are.

United Kingdom: Overview of Regulatory Landscape

Though there is no defined regulation yet, the UK aims to establish a flexible regime that would regulate the sectorial applications of AI, rather than the underlining software and systems.

The UK government first published a White Paper in March 2023, then updated in August, detailing the country's plan to implement a guideline for regulating AI. This is still very much work-in-progress, and no draft legislation has been published yet.

- AI applications and development are currently regulated by existing set of rules and laws. The current regulatory framework raises concerns across industries over the lack of a unified set of rules.
- The aim of the new regulation is to lead the international landscape on AI governance with a pragmatic and proportional regulatory approach, providing a clear, pro-innovation regulatory environment for foundational AI companies and systems. The regulatory framework will be based on five principles: Safety, security and robustness; Appropriate transparency and explainability; Fairness; Accountability and governance; Contestability and redress.
- The scope of the regulation will focus on establishing a regulatory framework based on context, outcomes and the use of AI, rather than the technology itself.
- It will not put immediately into enforce legislation, as it is believed to hinder innovation in the sector. The regulation will be first issued on a non-statutory basis.

UK AI Safety Summit

The UK hosted the first global AI Safety Summit in November 2023, where the main topic will be how international co-operation in regulating AI systems can help mitigate risk and will see both global leaders and AI companies attending.

- The summit also has geopolitical implications: the UK can use its position outside the EU to promote its more flexible, less strict approach on AI, also considering that several EU Member States attended the summit.
- UK Government officials are also currently in negotiations with large AI system developers such as DeepMind for permission to examine their LLMs, which would be unprecedented, as they seek to understand more in depth the technology.

Trajectory of AI Regulation

Current regulation on AI applications in UK is viewed as too fragmented. The government is structuring a flexible regulatory framework to regulate AI use and development by use cases and industry, rather than the software behind.

Current Status:

- Scattered current rules on AI use and development criticized by industry as hindering innovation.
- UK is developing a new regulatory framework, use case-based, which will not be enforced straight away.
- The British government is trying to establish the country's leading position for a flexible environment for AI developers and users.

Future Status:

- Official regulation still not on the table. The drafted white paper provides an overview on the approach, which is positively seen by industries.
- UK is currently looking not to ban the use of biometric recognition technologies, opposite of the EU. But privacy concerns are rising.
- The November Summit will serve as a starting point for the UK to start drafting an official regulatory framework, which however will not be legally binding at first.

China: Overview of AI Regulatory Landscape

China is implementing some of the strictest AI laws around the globe. In contrast to the EU, however, China is focusing primarily on specific applications and classifying risk based more on national security.

- **Administrative Measures for Generative AI Intelligence Services (2023, published):** Rules that came into effect on August 15, 2023 that regulate generative AI services. Regulations include requirement for licenses to operate, regular security assessments, and adhering to Chinese values.
- **Administration of Deep Synthesis of Information Services (2022, published):** Governs “deep synthesis technology” (AI/ML, algorithmic processing systems). It stipulates, among others, that users must consent to their image being used in outputs, disclosure of the use of deepfakes via labels, and prohibition of content that endangers national security and interests.
- **Internet Information Service Algorithmic Recommendation Management Provisions (2022, published):** This law regulates algorithms in apps and websites used in China. Key provisions range from requiring an online database of algorithms with “public opinion properties” or “social mobilization capabilities,” requiring regular reports, and transparency surrounding how the algorithm works, including its database.
- **Data Security Law (2021, published):** This law requires data localization of any data on Chinese citizens, meaning that all domestic and foreign enterprises alike must be housed in mainland China. Export of data is forbidden without a “technology review.”
- **Cybersecurity Law (2017, published):** In the context of AI software, this law has a number of strict provisions around data. For example, Article 37 also requires that all data be stored in servers in mainland China. There are other strict security obligations of “network operators,” which includes social media platforms, app developers, and tech companies.
- **Emphasis on AI Categories:** Unlike the EU, China does not broadly define AI based on a risk system. Its current legislation rather defines AI based on specific application category or technology, such as generative AI versus algorithms versus deep learning. This division is possible due to the structure of the government allowing for speed of legislation and implementation. Legislation is therefore fast, responsive, and vertical.
- **Security and Political Concerns:** China emphasizes security issues as a key point of regulation of AI and draws upon many previous laws in order to ensure AI systems do not pose any perceived threat to security. Given the lack of foreseeability and occasional “black box” nature of AI, however, this may prove a considerable challenge.

China’s regulations are of specific types of technology and algorithms and not necessarily tiered into risk categories. Legislation tends to be vertical – that is, it addresses problems as they arise. However, the rules being legislated are strict, and this is likely a long-term trend.

The Hiroshima Process

On October 30, 2023, members of the G7 issued a statement around commitment to ensuring both innovation and development in AI as well as proper human rights protections. They highlighted two documents that provide voluntary frameworks for AI ethics among G7 nations: the Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems and the Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems.

- **International Guiding Principles for Organizations Developing Advanced AI Systems:** Aims to promote safe, secure, and trustworthy AI worldwide and provides guidance for organizations developing and using the most advanced AI systems, including the most advanced foundation models and generative AI systems. Sets forth a non-exhaustive list of principles for organizations, including proper risk mitigation, transparency, and more.
- **International Code of Conduct for Organizations Developing Advanced AI Systems:** Aims to provide voluntary guidance to organizations by advocating actions such as taking appropriate measures for risk identification

Current Key Characteristics of China’s AI Regulation Regime

- **Innovation vs. Regulation:** Previously, China vigorously tackled AI and algorithms among other technology-related sectors. Now, however, China finds itself in a balancing act of courting private sector AI investment and maintaining the established order, targeting its sagging economic growth in particular. This means that there may be some limited opportunities for businesses in this field, but they are subject to stringent requirements.

and mitigation, public disclosure of AI capabilities, measures for data protection, and more.

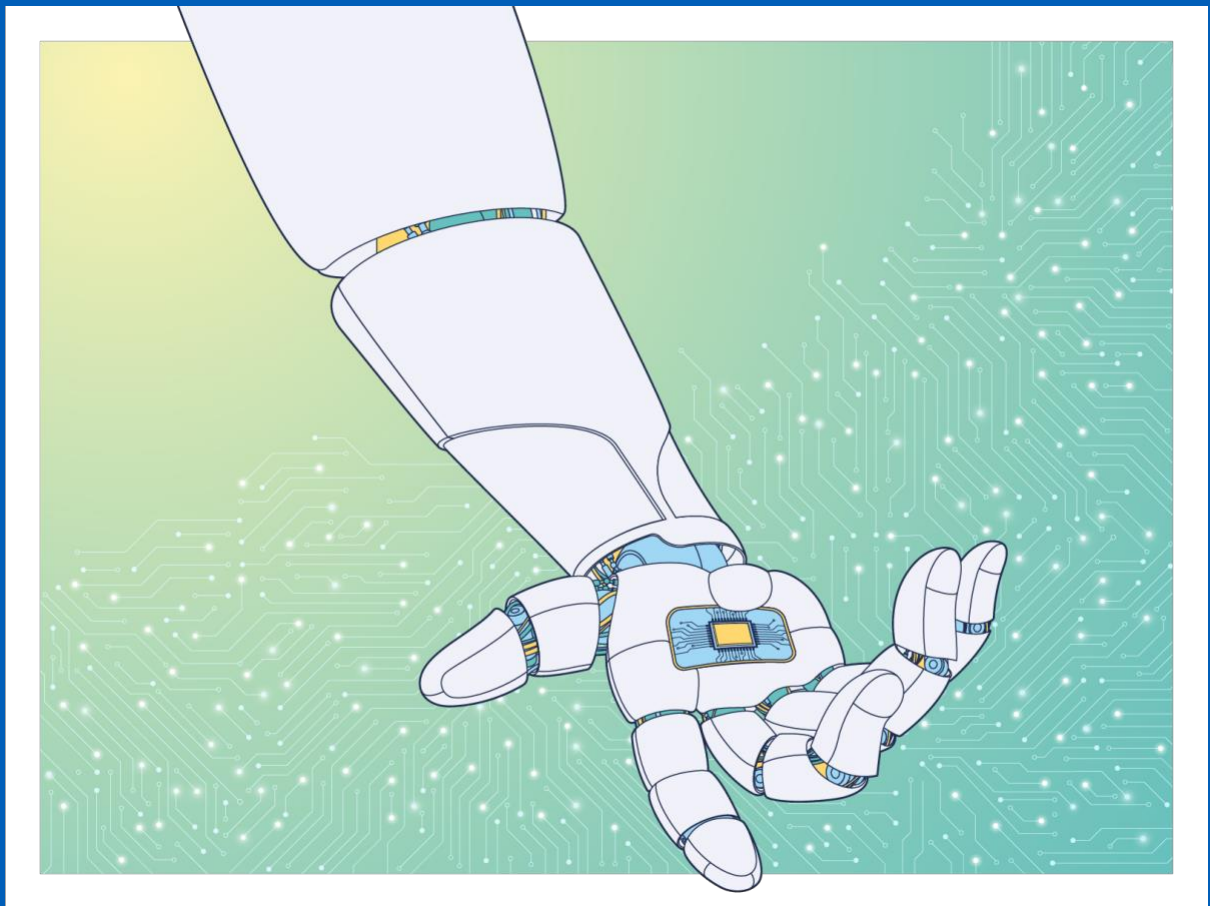
- **Significance and Challenges:** These two documents are non-binding international guides and thus do not have any enforceability or accountability mechanisms. They rely on the volition of companies to implement these principles. However, these frameworks also have the potential to influence trajectory of legislation in these G7 countries, most notably in Japan which currently has none. G7 economies are also highly influential and this legislation can set the stage for other economies to follow.

References

- [1] IDC's The Generative AI Arc Survey, August 2023.
- [2] IDC's European Emerging Technologies Survey September 2023.
- [3] IDC's AI Strategies View Survey, December 2022.
- [4] IDC's AI Tracker Forecast 2023H2, October 2023.
- [5] IDC's AI Spending Guide Forecast 2023H2, October 2023.

Jack Vernon is a senior analyst at IDC.

New hardware





Semiconductors are essential to today's societies and they underpin the Next Computing Paradigm. Their importance has been recognized at the highest political levels, resulting in the passing of the U.S. and European Chips Acts and similar instruments across the world. Investments made over the next few years will be critical to European competitiveness and security.

New hardware

by Paul Carpenter, André Hebben, Paolo Azzoni

Three key factors make this an exciting time for hardware in Europe. Semiconductor technologies have been, for a long time, critical to competitiveness, prosperity, and security. Their importance has now been recognized at the highest political levels, resulting in the recent enactments of Chips Acts across the world, leading to significant investments in fundamental hardware technologies. At the same time, the ending of Dennard scaling and slowing of Moore's law, due to fundamental physical limits, are bringing a diversity of system architectures from heterogeneous and specialized accelerators, to new memory types, to processing in memory, network and storage. This leads to, and is supported by, the rise of open-source hardware development, which reduces the barrier to entry for hardware development.

The largest development in Europe related to computer hardware, over the past year, has been the passing of the European Chips Act, in September 2023 [Azzoni23, diPaola23]. It promises to reduce Europe's dependence on non-EU suppliers and grow Europe's share of the global microchips market from 10% to 20% in 2030. The European Chips Act has three pillars. The first is the **Chips for Europe Initiative**, which aims to bolster Europe's technological capabilities and autonomy in design and production. It will be supported by a public investment of €4.175 billion,¹ which, together with private contributions, gives a total

investment of about €11 billion. This initiative will be implemented by the new CHIPS Joint Undertaking (JU), which replaces the former Key Digital Technologies (KDT) JU. The first CHIPS JU call for pilot lines has been opened, focusing on FD-SOI, leading edge technology at ≤ 2 nm, advanced packaging and wide bandgap semiconductors.

The second pillar facilitates exemptions from the EU's limits on State aid for "first-of-a-kind" fabs ("Integrated Production Facilities") and foundries ("Open EU foundries"). The aim is to attract investments in European semiconductor manufacturing, advanced packaging, test, and assembly, thereby leading to improved security of supply and resilience of the EU's semiconductor sector. The third pillar constitutes measures to **monitor the supply chain** and **intervene in supply-chain crises**, by building a coordination mechanism among Member States and the Commission to strengthen collaboration to report disruptions in the supply chain.

Similar initiatives are being developed in other countries or regions, including the U.S. Chips Act, the Chips 4 Alliance, and the initiatives of Japan, Korea and Taiwan. In this context, the European Chips Act stands out with its comprehensive approach, embracing scientific research, development of pilot lines, establishment of unique facili-

¹ €3.3B from the Chips for Europe Initiative plus €1.3B for the KDT JU.

ties, a new platform for chip design and manufacturing, competence centers to tackle the skills shortage, and strategies to monitor and prevent future chip scarcities.

With this important background, this section has three articles:

- **Qubits and pieces: Developments and Perspectives in Quantum Computing.** Recent advances in quantum technologies have attracted significant industrial attention, and countries have already invested vast resources to develop their capabilities. This article identifies the key components to build a powerful quantum ecosystem.
- **Open hardware and RISC-V: an exciting opportunity for Europe:** Open source is well established in software, as a global framework for collaboration and innovation, which benefits the whole landscape from HPC to cloud to IoT and AI. This article recommends investments in open source hardware, with RISC-V as the open source solution of choice.
- **Here to stay: specialized and heterogeneous computing:** Specialized hardware can improve performance and/or energy efficiency, since no architecture is perfect for all use cases. This article discusses the rise of specialized and heterogeneous hardware, due to the end of Dennard scaling and the rise of open source hardware.

References

[Azzoni23] Paolo Azzoni, Jean-Luc di Paola Galloni, and Dimitrios Serpanos. Chips Shortage and Chips Acts: The Inside View. IEEE Computer, November 2023.

[diPaola23] An industrial powerhouse in the markets of the future. Inside Magazine, November 2023—Issue 05. <https://intranet.inside-association.eu/publication/download/inside-magazine-5.pdf>

[Leiserson20] Charles E. Leiserson et al. There’s plenty of room at the Top: What Will drive computer performance after Moore’s law? Science **368**, 1079 (2020).

[Herz23] Douglas Herz. A Century of Moore’s law. <https://www.semianalysis.com/p/a-century-of-moores-law>

Paul Carpenter is a researcher in the Computer Sciences Department at Barcelona Supercomputing Center, Spain.

André Hebben is IT Solutions Manager at INSIDE Industry Association.

Paolo Azzoni is secretary general of INSIDE Industry Association.



The end of Dennard scaling is leading to a Cambrian explosion of specialized hardware, with better performance and energy efficiency, and optimized non-functional requirements. But it comes at the cost of programming complexity, difficulty in code reuse, and vendor lock-in.

Here to stay: specialized and heterogeneous computing

by Jaume Abella, Leonidas Kosmidis, Thomas Hoberg and Paul Carpenter

Until the early 2000s, while specialized approaches could always improve performance and/or energy efficiency, general-purpose processors had lower cost due to much higher volumes, as well as being more flexible and easier to use. Any niche special purpose chip was soon overtaken by the relentless advance of Moore's law and Dennard scaling. There was space for a small number of architectures: general purpose CPUs in PC and Macs (x86) and phones (Arm), graphics processing units (GPUs) for gaming and the film industry, and digital signal processors (DSPs) for audio, video and communications.

The release of Intel's first multi-core CPU in 2005, following the failure of its planned 10 GHz Tejas CPU, was a major change in direction for Intel and a significant milestone in the industry. For the next few years, much of the improved performance came from greater numbers of cores, but this was curtailed by limits on how much software can benefit from parallelism (Amdahl's Law), as well as practical constraints on the scaling of shared hardware structures.

In 2007, Nvidia released its first CUDA-capable GPU, having realized that its highly multi-threaded architecture, originally developed for 3D graphics rendering, could deliver improved performance for many other data-intensive compute tasks. The downside was that the key computational kernels and supporting code needed to use

Nvidia's new CUDA programming model, which exposed the key characteristics of GPU architecture to the programmer. Since GPUs delivered so much more performance than CPUs, many applications were adapted to use CUDA, spurred by significant investment from Nvidia and heightened interest in the community. These applications employ CUDA where it makes sense, while still using CPUs for the rest of the application. Examples that benefit include desktop image and video processing, high performance computing and artificial intelligence.

In the seventeen years since then, the rise of specialized and heterogeneous architectures is well known. Microsoft started using FPGAs for search. Amazon, Meta and Google have developed specialized chips for their applications. All top ten supercomputers except one have GPU accelerators. There are now hundreds of startups with AI chips. There is still an insatiable demand for more performance, due to the demands of the new computing paradigm, autonomous driving, LLM training and other applications. Meanwhile, the rise of open source hardware has greatly lowered the barrier to entry for the development of custom hardware (see article). Modularity theory (Carliss Y. Baldwin and Kim B. Clark) predicts that as competitive environments mature and technology becomes good enough, the performance advantages of specialized and integrated systems eventually lose ground to the cost and

compatibility advantages of a modular and standardized approach. But there are no signs that this is likely to happen again for computing systems in the foreseeable future. Computing systems are likely to be diverse and complex for some time to come.

Key insights

- Hardware that is specialized to the application can achieve **higher efficiency**, e.g., performance-per-watt or joules-per-operation, compared with a general-purpose architecture.
- For the foreseeable future, future computing systems will continue to be **heterogeneous**, driven by application needs and the **end of Dennard scaling**, and disruptive technologies (e.g., processing-in-memory, quantum computing, and neuromorphic).
- **Programmability** and software **portability** are of critical importance. There is as yet no **performance portable** programming environment that allows a single source codebase to achieve good performance across a range of architectures. There may never be one, but more progress towards this goal is needed.
- **Heterogeneity affects all levels of the software stack**: system architecture, operating system, middleware, applications.
- Heterogeneous architectures increase complexity, making **performance analysis, correctness testing** and **debugging** very difficult. Errors in the developer's source code may lead to obscure low-level errors in unfamiliar libraries. Similarly a seemingly trivial change may lead to a dramatic difference in performance, for obscure reasons.
- Independent application developers are **reluctant to adopt niche programming environments** unless the advantages are overwhelming. This leads to a “survival of the biggest”: CUDA, OpenMP, etc. New programming approaches should be **standardized to drive adoption**.
- If architectural diversity reduces investment and evolution of general purpose CPUs, more and more workloads may need to move to the cloud to get access to the right accelerators.
- Vertical integration risks increasing the market dominance of the largest technology companies, none of which are European.

Key recommendations

- Embedded hardware acceleration must be aware of user requirements and software limitations, e.g., accuracy, time-to-solution deadlines (inference speed) and maximum power consumption.
- Specialized hardware should be optimized for application semantic precision rather than bit-level precision, trading off arithmetic precision, operation accuracy or approximation, voltage scaling, reliability and others.
- Vendor lock-in should be prevented, especially since the vendors are not European.
- Interoperability and **composability** of hardware and software needs to be maintained.

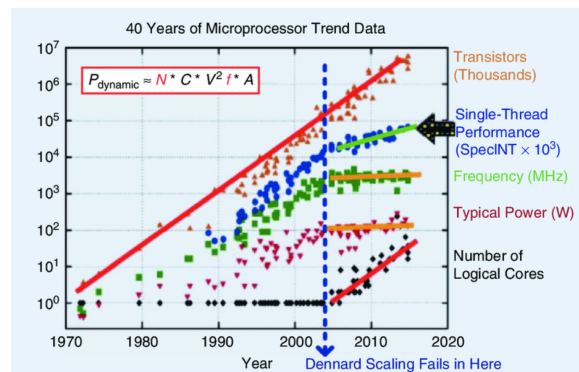


Figure 1 Moore's law continues to deliver increasing numbers of transistors, but Dennard scaling failed in the early 2000s. The consequences include greater numbers of cores. The plot was originally collected in 2010 for Horowitz et al, and variants have appeared in hundreds of papers and articles.

Survey of specialization and heterogeneity

Here we outline the main kinds of specialization and heterogeneity:

Graphics Processing Units (GPUs) are the most well-known type of accelerator and are specialized to handle highly parallel tasks. Unlike traditional CPUs that are designed to perform a wide range of computing tasks, GPUs are adapted for concurrency, making them exceptionally efficient for graphics rendering and data-heavy tasks in fields like scientific computing, machine learning, and video processing. The architecture of a GPU is tailored for multi-threading, achieved by housing hundreds or thousands of smaller and more

efficient cores. Additionally, GPUs are equipped with high-bandwidth memory, which allows for faster data transfer rates, enabling them to swiftly handle large datasets and complex calculations.

Nvidia GPUs are commonly programmed using CUDA, a parallel computing platform and programming model developed by Nvidia. CUDA allows developers to write software that can take full advantage of the unique architecture, by offering direct access to the GPU's virtual instruction set and parallel computational elements, enabling highly efficient processing for complex computational problems. However, CUDA is proprietary and exclusive to Nvidia GPUs. This exclusivity means that while CUDA is powerful, mature and widely used, it isn't compatible with GPUs from other manufacturers like AMD or Intel. Another example is AMD's ROCm (Radeon Open Compute), an open-source platform designed specifically for AMD GPUs. In order to avoid fragmentation and enable interoperability, Khronos introduced open APIs such as OpenCL and Vulkan, which are lower level than CUDA and therefore allow for more control in expense of programmability, as well as a higher level C++ based solution, SYCL. All major GPU vendors participate in Khronos and co-operate in the definition and standardization of its APIs.

Neural Processing Engines (NPEs) are specialized hardware designed to efficiently handle the computations required for neural networks, which are at the core of modern artificial intelligence, image and machine learning applications. Unlike general-purpose CPUs, NPEs are optimized for the parallel processing and matrix operations that are prevalent in deep learning tasks, offering faster performance and greater energy efficiency for AI computations. Like GPUs, they usually have a high number of processing cores and high-bandwidth memory interfaces, but their architecture is tailored for tasks like convolution and matrix multiplication, including at reduced precision.

Programming these engines is usually done using high-level frameworks and libraries tailored to AI and ML, such as TensorFlow, PyTorch, or Keras. These frameworks abstract the complexity of directly programming the hardware, allowing developers to focus on designing and training neural networks using high-level, pythonic interfaces. Underneath these frameworks, lower-level libraries and drivers specifically optimized for NPEs translate these high-level instructions into efficient

hardware operations. Additionally, some NPEs support automatic optimization where the engine itself can dynamically adjust computing resources based on the workload, making them highly efficient for varying AI tasks. This combination of specialized hardware and sophisticated software stacks makes NPEs a cornerstone in advancing AI and ML technologies.

Field Programmable Gate Arrays (FPGAs) were introduced by Xilinx in 1985, aimed at rapid prototyping and the implementation of custom digital logic in low-volume applications. They were first recognised as compute accelerators in the 2000s, and there was a significant expansion in the late 2010s, partly spurred by Microsoft's Cata-pult project, which employed FPGAs to accelerate web search and real-time AI. In 2024, FPGAs are still a niche device for several reasons. First, FPGAs have a spatial architecture, which is unfamiliar to software developers, since the algorithm is laid out in space in the same way as hardware. High-level Synthesis (HLS) greatly improves matters, since software developers can use extensions of familiar programming languages such as C and C++, but specific annotations and custom control of data transfers are required. Second, the mapping of an algorithm to the hardware depends on the FPGA size and its numbers of resources across multiple dimensions (number of logic cells, DSP slices, amount of embedded SRAM, number of QSFP+ ports, etc.). If the design doesn't fit, then it doesn't compile. Thirdly, the software ecosystem for FPGAs is still relatively immature and varies significantly between the two major U.S.-based vendors. Moreover, achieving performance that matches or exceeds that of GPUs or more specialized accelerators is a complex and challenging task.

Vector Processing Units (VPU)s have been used for decades in order to provide high performance processing. First employed by the ILLIAC IV supercomputer in the 1960s, vector architectures were the fundamental design of all later supercomputers. Vector processing comes in many different variants ranging from very long vectors to short Single Instruction Multiple Data (SIMD) units. A vector processing unit can be either tightly integrated in an in-order, out-of-order and/or multi-threaded processor or operated as a decoupled accelerator such as a GPU. Moreover, it can operate on a distinct register file or reuse existing register files, a concept known as SWAR (SIMD within a Register) or packed SIMD. Different design choices provide different tradeoffs in terms of

area, performance and latency. Thanks to Moore’s Law, in the 1990s vector processing units found their way into general purpose processors in the form of SIMD instructions, mainly for the acceleration of graphics and multimedia tasks. Naturally, embedded processors followed, and with the passing of the years processor designers included longer and longer vector lanes and registers. In the last decade, vector length agnostic ISAs were proposed, first by Arm and later by the RISC-V foundation, allowing software portability among architectures with different vector lane width. Moreover, significant progress has been performed in vectorising compilers but not all compiler backends have the same maturity. Consequently, still there is no performance portability among the continuum of vector architectures, and therefore more research on high-level languages is required. Similar to the rest of the hardware technologies, most of the vector processors nowadays are not European. Within the European Processor Initiative (EPI), significant advances have been made with the design of several vector processing units which explore the strengths and the weaknesses of each of their design choices.

Video coding chips, also known as video encoding or transcoding processors, are specialized hardware used by companies like Google YouTube and other streaming service providers to efficiently process and stream video content. 500 hours of content is uploaded to YouTube each minute, and this needs transcoding to every viewable resolution in H264; popular videos are also transcoded using a more complex but more space-efficient codec such as VP9.

Neuromorphic Computing (NMC) was first developed by Carver Mead in the late 1980s. It describes the use of large-scale adaptive analog systems to mimic the human nervous system. Originally, the approach was to use elementary physical components of integrated electronic devices (transistors, capacitors, ...) as computational primitives. In recent times, the term neuromorphic has also been used to describe analog, digital, and mixed-mode analog/digital hardware and software systems that transfer aspects of structure and function from biological substrates to electronic circuits (for perception, motor control, or multi-sensory integration). Today, the majority of NMC

implementations are based on CMOS technology. Interesting alternatives are, for example, oxide-based memristors, spintronics, or nanotubes.

Memory systems The basic architecture of a stored program computer still follows the von Neumann architecture outlined in 1945. Commodity memory is built in high volume and generally optimized for increasing capacities not speed. In 1995, Wulf and McKee coined the term “memory wall”, pointing out that there is a limit on how much caches can mitigate the growing gap between processor and memory speeds. Recently, a number of new device types (PCM, STT-RAM, etc.) and memory interfaces (HMC, HBM) have started to offer diverse tradeoffs among capacity, latency, bandwidth, volatility and energy consumption. A traditional memory hierarchy (registers, cache, DRAM and disk) is making way to a collection of memory devices on an equal ranking. This is facilitated by the advent of Compute Express Link (CXL), which promises exciting times for memory systems through a standardized high-bandwidth and cache coherent interface. Data placement and migration become more difficult, and they can be done in various places, e.g. in hardware (e.g. now-discontinued Optane in memory mode), transparently by the OS (at page granularity) or by the runtime system or support library (at page or object granularity), or even by the compiler. If not, the whole zoo of memory types may simply be exposed to the application.

Processing in memory is a general term that covers any form of integration of processing capability into the memory system.¹ It was initially proposed in the 1970s,² but it has received renewed interest as a way to avoid the memory wall. There are several approaches for processing in memory, including computation inside the memory array or peripheral circuits, as well as logic layers and dies integrated alongside memory dies within the silicon package. Examples include Micron’s Automata Processor, which adapts the DRAM memory arrays and peripheral circuits for finite automata, UPMEM’s DRAM Processing Units (DPUs), which are embedded next to the DRAM bank, and Samsung’s Function-In-Memory DRAM (FIM-DRAM, which) integrates SIMD processors in the memory bank. The multiplicity of approaches represents challenges for OS management (data

¹ ETP4HPC, https://www.etp4hpc.eu/pujades/files/ETP4HPC_WP_Processing-In-Memory_FINAL.pdf

² H. S. Stone, “A Logic-in-Memory Computer,” IEEE Transactions on Computers, vol. 19, 1970.

placement, security, resilience) and the programming environment (translating algorithms to the architecture, as well as debugging and performance analysis).

In-network computing refers to the integration of computational capabilities directly into the network, specifically within network switches and routers of a data centre or HPC system. This approach contrasts with traditional network architectures, where data processing is performed solely at the endpoints, such as servers or client devices. The key idea is to leverage the network hardware itself for certain types of data processing tasks. By doing so, it can reduce the data processing load on servers, decrease network latency, and improve overall system efficiency. This is particularly beneficial for applications involving large-scale data movement, real-time analytics, and distributed computing scenarios where reducing latency and data movement is crucial. In this architecture, network devices are equipped with processing units that can perform a range of functions like data aggregation, filtering, or even more complex operations like machine learning inference, directly on the data as it passes through the network. In-network computing is an emerging field, driven by advancements in networking hardware and the increasing demands of modern applications for faster data processing and reduced latency. It represents a shift in how data networks are traditionally viewed, evolving from mere data conduits to active processing nodes within an IT infrastructure.

Quantum computing was first suggested by Richard Feynmann in 1981 for the simulation of quantum systems. Revolutionary algorithms for prime factorization (Shor's algorithm, 1994) and unstructured search (Grover's algorithm, 1996) showed that quantum computing has the potential for exponential speedup for certain algorithms over classical computers. A quantum computer is likely to appear as another kind of accelerator in a heterogeneous system, so quantum computing is related to the topics of this paper. Recent advances have led to impressive quantum computers, taking a number of different approaches. More details are in the article on quantum computing.



Figure 2 A “Cambrian explosion” in computer architectures, showcasing an array of diverse and innovative chip designs (made by ChatGPT).

The programmability challenge

The move from general purpose processors to a diverse set of specialized and heterogeneous systems improves energy efficiency and performance. But it raises software complexity, especially because a single application needs to run well across a range of systems. A key challenge will be to identify and support high-level software abstractions that exploit the full capabilities of the hardware, while abstracting away most of the architectural diversity. The main application code should ideally be platform independent, with an open, clean, stable and dependable interface to the hardware. Previously, for performance-critical code on general purpose processors, this was single-threaded C or C++. But now this is under pressure.

Ideally, the platform-independent part should avoid direct references to specific resources and their availability, and instead provide declarative information that is sufficient to optimise behaviour. A number of different strategies, frameworks and libraries have been proposed, and are under investigation, but none are yet a general solution. OpenCL offers functional portability across heterogeneous platforms, including different types of GPUs and CPUs, but good performance requires the code to be optimized for the architecture. Performance portability, meaning that the same code achieves optimal performance on all platforms, and it is a holy grail in the field. SYCL has provided some promising results in this direction, but

its adoption is not very wide and neither all platforms have an optimized backend. Overall, more intelligence is needed in the programming environment, from compilers to runtime systems to performance and debugging tools. One avenue is to explore the use of artificial intelligence techniques, as exemplified by large language models such as ChatGPT-4, to perform the translation from a declarative specification of the problem into an optimized source code or intermediate representation. The progress in related techniques has been spectacular in recent years, so this is an interesting approach. However, special attention should be paid that standardized solutions with open specifications will be used, in order to allow interoperability within the heterogeneous environment and avoid vendor lock-in.

Domain Specific Languages (DSLs) are specialized programming languages designed for a particular application domain. They offer enhanced productivity and ease of use for domain experts, who can focus on domain-relevant aspects and abstract away the complexities of programming. They also reduce the likelihood of certain types of errors and bridge the gap between domain experts and developers, by using terminology and concepts familiar to practitioners in the field. While DSLs are user-friendly for domain experts, there may be a learning curve for developers not familiar with the domain. Also, they may suffer from limited support and community resources, unless they are widely adopted. Integrating DSLs with other systems and languages can also be complex, as they might not easily interact with software written in general-purpose languages. It is beneficial to avoid a proliferation of DSLs by identifying and using common reusable features, leading to a small number of more generic DSLs or DSL frameworks, and/or their integration into more general purpose programming environments. DSLs are most beneficial when the advantages in productivity and performance outweigh the costs and limitations in scope and support.

Task-based programming first appeared in the 1990s and it is already mainstream in high performance computing through its adoption in the OpenMP 3.0 standard in 2008.³ The tasking models separates the program into (a) a decomposition of a algorithm, in an architecture independent

way, into tasks with dependencies, (b) a runtime system, aware of the architecture, which manages scheduling, the choice of execution device and data locality, and (c) implementation of the low-level kernels, which is typically device-specific. It is especially beneficial for asynchronous parallelism and complex system architectures.

Finally, as the general-purpose CPU becomes relegated to a mainly orchestration role, management by the system software and middleware must not become the bottleneck. Such activity must be moved outside the critical path of computation and communication.

Conclusion

Heterogeneous and specialized architectures are not new. They have been consistently featured in the HiPEAC roadmap since the first “HiPEAC vision” in 2007. The end of Dennard scaling has led to a new golden age of computer architecture and a Cambrian explosion of approaches, from GPUs, to neural acceleration, to FPGAs, as well as the integration of processing capabilities into memory and storage. There is as yet no performance portable programming environment that allows a single source codebase to achieve good performance across a range of architectures. There may never be one, but more progress towards this goal is needed.

³ https://www.etp4hpc.eu/pujades/files/ETP4HPC_WP_Task-based-PP_FINAL.pdf

References

- [1] Stanford University, "Genie: the Open, Privacy-Preserving Virtual Assistant," Open Virtual Assistant Lab at Stanford University, [Online]. Available: <https://genie.stanford.edu/>. [Accessed 29 November 2022].
- [2] T. Vardanega and M. Duranton, "'Guardian Angels' to protect and orchestrate cyber life," *HiPEAC Vision 2021*, pp. 48-53, <https://doi.org/10.5281/zenodo.4719375>, January 2021.
- [3] T. Vardanega, M. Duranton, K. De Bosschere and H. Munk, "Past, present and future of the web: A HiPEAC Vision," 2019. [Online]. Available: <https://www.hipeac.net/media/public/files/46/7/HiPEAC-2019-Comic-Book.pdf>.
- [4] "Stable Diffusion on GitHub," [Online]. Available: <https://github.com/CompVis/stable-diffusion>. [Accessed 24 November 2022].
- [5] "WebAssembly," [Online]. Available: <https://webassembly.org>.
- [6] "Ballerina," [Online]. Available: <https://ballerina.io>.
- [7] R. P. Feynman, "Simulating physics with computers," *Int. J. Theor. Physics*, pp. 21, 467-488, 1982.
- [8] J. Preskill, "Quantum Computing in the NISQ era and beyond," *Quantum*, p. 79, 2018.
- [9] "NUKIB - Tsjech National Cyber and Information Security Agency," [Online]. Available: <https://nukib.cz/cs/infoservis/aktuality/1984-nukib-pripravil-podpurne-materialy-pro-ochranu-pred-hrozbou-v-podobekvantovych-pocitacu/>.
- [10] T. Attema, J. Diogo Duarte, V. Dunning, M. Lequesne, W. van der Schoot and M. Stevens, "Het PQC-migratie handboek (in Dutch)," 2023.
- [11] "Post-Quantum Cryptography - Setting the Future Security Standards," [Online]. Available: <https://www.nxp.com/applications/enabling-technologies/security/post-quantum-cryptography:POST-QUANTUM-CRYPTOGRAPHY>.
- [12] A. Singh, K. Dev, H. Siljak, H. Joshi and M. Magarini, "Quantum Internet—Applications, Functionalities, Enabling Technologies, Challenges, and Research Directions," *IEEE Communications Surveys & Tutorials*, pp. 2218-2247, 2021.
- [13] B. Kantsepolsky, I. Aviv, R. Weitzfeld and E. Bordo, "Exploring Quantum Sensing Potential for Systems Applications," *IEEE Access*, pp. 31569-31582, 2023.

Jaume Abella is a researcher in the Computer Sciences Department at Barcelona Supercomputing Center, Spain.

Leonidas Kosmidis is a researcher in the Computer Sciences Department at Barcelona Supercomputing Center, Spain and serves as an Outreach Officer for Khronos' SYCL SC (Safety Critical) WG.

Thomas Hoberg is technical director at Worldline Labs, Frankfurt, Germany.

Paul Carpenter is a researcher in the Computer Sciences Department at Barcelona Supercomputing Center, Spain.



Efforts to build an operational quantum computer are steadily progressing, but the “going is tough”. And it is not just a matter of battling noise in qubits: it is about building a complete industry, including fabricating qubits, integration into computer systems, incorporation into HPC, and quantum computer programming. As well as efforts in plain physical systems to get quantum computers up and running.

Qubits and pieces: Developments and Perspectives in Quantum Computing

by Artur Garcia-Saez, Christian Gamrat, Harm Munk and Paul Carpenter

Quantum technologies are becoming a revolution due to recent technological advances that have attracted a lot of industrial attention. Among these technologies, Quantum computation plays a central role due to its potential to provide a novel source of computational power to solve very hard problems. These results will be transformed into a novel set of devices of high precision, with the promise of novel applications and capabilities. Their development requires an environment not only of scientific research centers, but also auxiliary industrial partners with a high level of specialization.

Due to the strategic potential of this novel technology, many countries have already invested vast resources to develop their industrial capabilities, and to attract scientific talent. Europe has already started an initial set of projects defining its strategy for the upcoming years, but a careful analysis of the requirements to develop a global industry among different countries is required. The result of this analysis identifies the key components to create a powerful ecosystem for the upcoming years.

Key insights

- The life cycle of some of the pioneering companies in QC is already showing stress symptoms, as the market conditions have evolved since the initial appearance of private actors. These changes may condition the future of the field, as companies have played an important role in recent developments.
- Europe has pioneered several qubit technological paths: superconducting qubits, trapped ion qubits and neutral atoms, and is very active in the development of those technologies.
- The **race for making systems out of qubits** is currently being led by players in the United States but Europe is narrowing the gap thanks to a number of innovative solutions developed by start-up companies. A successful European quantum computer industry also requires (re)building its industry providing the physical infrastructure required for quantum computing hardware.
- Most available quantum programming frameworks originate from major US companies. European players are entering the discipline

and are already proposing approaches for experimenting with quantum programming. Europe as a whole (European Commission and member states) is a strong competitor.

- Europe is competing to train, attract and retain quantum talent. Education and training are critical for the European quantum computing ecosystem to be able to get up to speed. Access to quantum hardware and top researchers are essential to attract and retain talent.
- The quantum computer will be part of a hybrid quantum/classical computer. A good **synergy between quantum and computer science communities** is required at both the hardware and software levels. The Quantum and HPC communities need to strengthen their cooperation and learn from each other.

Key recommendations

- Support and advance research in the **system architecture and software stack** for quantum computing. With European assets in system, software and HPC technologies, this is a significant opportunity for Europe.
- Invest in the development of **design tools and libraries** for quantum chip design and in their integration with classical technologies so as to provide an easier path towards prototyping and fabrication for both research and industry.
- The **chip technologies** and industries will play a key role in developing and maturing new design methods, fabrication techniques and novel packaging for quantum computers. Recognize that they need to be supported and, in this context, the role of quantum pilot-lines will be key to provide access to prototype systems to the whole community including SMEs.
- Design support and investment programmes to stimulate and promote the introduction of quantum technology in industry and society. Supporting technologies at large, e.g. infrastructures, should also be reinforced in order to build on the core quantum technologies (e.g. qubits) to make the advances in quantum technology real.
- Since a quantum computer will be a hybrid classical/quantum system, high-performance computing (**HPC**) **infrastructures** will play a key role in integrating the required **hybrid quantum system stack**. Develop the integration of quantum accelerators into future exascale infrastructures. Promote the emergence of a European quantum cloud connected to quantum-hybrid HPC platforms. Develop software aspect, quantum stack. Hybrid computing: the marriage between HPC and Quantum.
- Position Europe as a key player in the global race: Europe is the leader in the foundations of quantum physics: it should keep promoting basic research and novel ideas.
- Europe needs to increase its support for new and small enterprises in the field, as well as promoting funding for scaling up, rather than being acquired by a non-European company. Developing and sustaining an ecosystem for these startups is also of paramount importance to prevent fragmentation.
- Europe should promote academic and industrial research centers independent of U.S. and Chinese companies. Larger companies should be encouraged to take risks on quantum research.
- Europe should also move to practical quantum computation, including the consolidation of a small number of easily accessible, highly usable quantum computing sandboxes in the cloud, with a strong eye towards ease of use.
- Promote an attractive European quantum computing ecosystem with a balanced salary/benefits structure that will encourage talented people to stay in or relocate to Europe.
- Many aspects of quantum computing and its applications are under development and require strong cooperation between the classical and the quantum computing communities: the communities are in a perfect position to learn from each other and thus strengthen their position in the world without reinventing the wheel.
- In a competitive and cutting edge technological industry which quantum computing is, efficient and effective coordination among and across nations is very important.
- Europe as a whole (European Commission and member states) is a strong competitor when levels of public funding are taken into account, but probably needs improved coordination to more effectively leverage available funding.

Quantum computing primer

Any computation makes use of some physical process that will transform data according to some rules. If we can manage to control these rules and make them equivalent to some logical operations, then we can make a computing device. Up to now, the rules used to build computers have been based on classical phenomena: the set of physical theories developed until the early 20th century.

Quantum theory was born to explain a number of experiments related to matter and light at the small scale that could not be explained by classical physics. As a side product, the formulation of this new physical theory provided a new set of rules that generalize those of classical logics. Using these rules, we can build a novel type of computing device based on the interactions described by quantum theory.

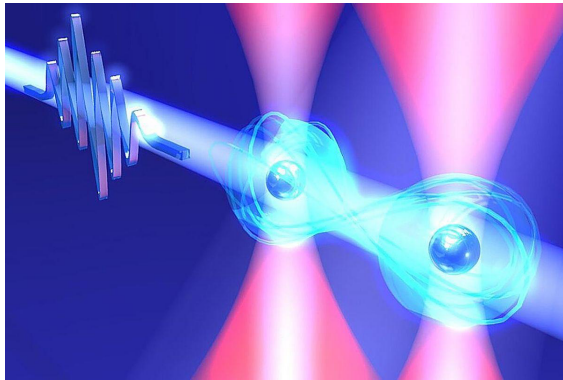


Figure 1 Illustration of two qubits in blue light with red light vertically (source: <https://spectrum.ieee.org/media-library/illustration-of-two-qubits-in-blue-light-with-red-light-vertically.jpg?id=30824034&width=1200&height=804>)

Developing algorithms for this new type of computer has provided a novel perspective on the theory of computation and provided a very powerful result: by using quantum operations, one can design algorithms that solve problems (exponentially) faster than conventional computers. This is a major result, as advances in speed were based on improvements of the algorithms or improvements of the hardware, such as increasing the speed of a processor. However, the dramatic speedup for some problems using quantum computers are a result of the combination of a new formulation and a new set of operations, not available in classical logic. Quantum computation, a general case of the computers we use daily, provides more powerful computing devices. Shor's algorithm to factor a number in its prime factors is often

quoted as an example of this speedup, because it opens the road to find cryptographic keys. Once quantum computers become practical, this may well pose a threat to secure, encrypted communication. The key point here is that this requires a significantly larger quantum computer than is available now: for factoring a key of several thousand bits a logical qubit register of several thousand qubits is required. With the current state-of-the-art, this requires millions of physical (noisy) qubits. Such a large quantum computer is at least 10 to 15 years away, if not more (Preskill, 2018).

Nevertheless, quantum computing is casting its shadow ahead. And governments and organisations are already taking countermeasures. The government of the USA, e.g., now requires all its services to prepare for the security threat quantum computing will pose. It means that current encryption technologies are to be upgraded to a degree that even challenges quantum computing.

The European Union is preparing for this post-quantum era as well, mostly in the form of recommendations. Several national initiatives have evolved in the recent past. (See (NUKIB - Tsjech National Cyber and Information Security Agency, sd), and Attema, et al., 2023) (PQC=Post-Quantum Cryptography).

In recent years, the connection between HPC and quantum computing started taking shape. Quantum computers with their complicated support electronics are, from a computer systems architectural point of view, accelerators, or coprocessors, not stand alone systems. In 2021, work has begun on the EuroHPC JU "Pilot on quantum simulator" project <HPC|QS>. The aim of <HPC|QS> is to prepare European research, industry and society for the use and federal operation of QCS. <HPC|QS> is developing the programming platform for the QS, which is based on an emulating platform provided by a European company, and the deep, low-latency integration into modular HPC systems based on ParTec's European modular supercomputing concept. A twin pilot system of QSs, developed as a prototype by a European company, will be implemented and integrated at CEA/TGCC (France) and Forschungszentrum Jülich/Jülich Supercomputing Centre (Germany), both hosts to/of European Tier-0 HPC systems. The pre-exascale sites BSC (Spain) and CINECA (Italy) as well as the national Quantum Learning Platform at ICHEC (Ireland) will be connected to the TGCC and Jülich Supercomputing Centre via the European data infrastructure

FENIX (Ref QFLAG SRIA 2030, to be published). The <HPC|QS> serves as a seed project for the upcoming EuroHPC QS program, after the selection of 6 European centers as hosting entities for a new generation of Quantum computers that will operate the systems on behalf of the EuroHPC JU. This quantum computer infrastructure will support the development of a wide range of applications with industrial, scientific and societal relevance for Europe, adding new capabilities to the European supercomputer infrastructure.

It is often asked what the “killer application” of quantum computing will be. Without available practical quantum hardware allowing for large scale experimentation, that question is difficult to answer. For the near future, whatever that means for quantum computation, it seems to be the simulation of physical quantum systems, such as the structure of complex molecules. A slightly different quantum computational type of device, based on Rydberg atoms, is a likely candidate to provide the computational power for such simulations.

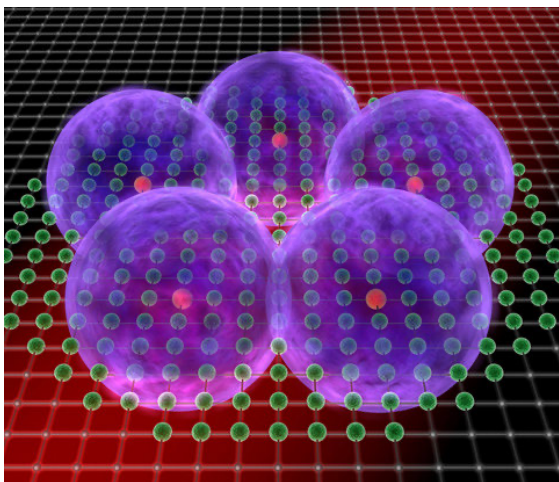


Figure 2 A pentagon of super atoms: The illustration depicts the densest possible ordering of five Rydberg excitations in an ensemble of rubidium atoms that are pinned in an optical lattice (source: <https://www.mpg.de/research/rydberg-excitation-quantum-gas>)

Computer science has developed a full theory to classify the difficulty of computational problems. Under this approach – termed *Computational Complexity*– some problems have been identified as easy to solve, while others may become hard to solve for a computer. This approach poses a limitation on the computational problems

we may solve, unless one finds better algorithms to solve these hard tasks.

The **Deutsch-Jozsa algorithm** is a good example of a Quantum algorithm that solves a problem classified as hard from the classical point of view. However, it becomes an easy task using a Quantum device running this algorithm. The problem, namely assessing the properties of a mathematical function of many variables, is hard classically because the function has to be evaluated many times, and is the number of repetitions that consumes the computational power. The Quantum approach uses a superposition of states to obtain the outcome using a single evaluation of the function.

This is a problem specifically designed to be easy for a Quantum computer but hard for a classical machine, although of little practical interest. However, by showing that Quantum computers can be exponentially faster than classical computers it paves the way to look after practical algorithms that may show a similar speedup over practical tasks.

The global race for quantum computing

Quantum computing is built on the foundations of quantum mechanics laid out by European scientists at the dawn of the 20th century. Since then, Europe has continued to excel in fundamental quantum physics. However, now that the utilization of quantum computing and communication are emerging we are entering an era of global competition in several very concrete fields.

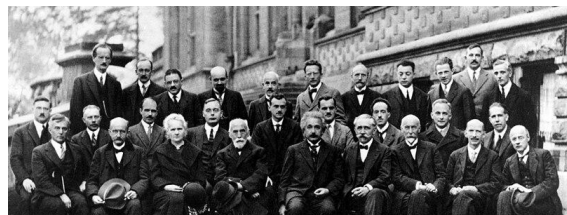


Figure 3: Solvay conference 1927

Europe pioneered several technological paths for qubit technology (first demonstration of superconducting quantum gates [11], trapped ion qubits [12]) and has a strong position in promising technologies such as neutral atoms (Pasqal: <https://www.pasqal.com/>) and semiconductors.

Europe must get up to speed in quantum computing systems hardware. Proofs of concept and innovative solutions are developed by pioneering European startup companies and research labs, but larger quantum chips so far have been made by US companies. Europe is now changing gears as suggested by the roadmaps from the "Quantum Flagship Initiative" (<https://qt.eu/>) and the association of European Industrial for Quantum technologies (<https://www.euroquic.org/>).

In software, most available quantum programming frameworks originate from major American companies. In Europe, some players (ATOS: <https://atos.net/en/lp/myqlm>, VeriQcloud: <https://veriqcloud.com/>, Quantum Inspire: <https://www.quantum-inspire.com/>) are well in the race and already propose approaches to allow experimenting with quantum programming using quantum computer simulators and quantum computer hardware. Also important to mention are the efforts of hardware start-ups that experiment with their platforms by providing cloud access to their programming tools like Quandela cloud (<https://www.quandela.com/cloud/>).

Europe is struggling with funding among nations. The overall public funding (EU as a whole plus individual member states) compares well by being head to head with the U.S. and China, but this hides the fact that each member state has its own agenda and funding strategies. Initiatives such as the already mentioned Quantum Flagship should bring more coordination. The situation is very different for private funding needed by the industry. Due to the lack of significant risk funding opportunities in Europe, companies might not find the resources to support their growth: some might find better opportunities elsewhere, some might simply die.

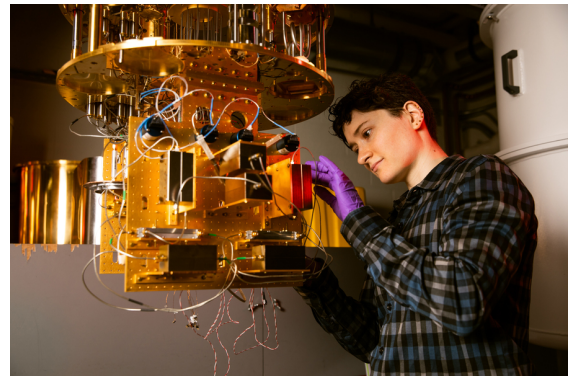
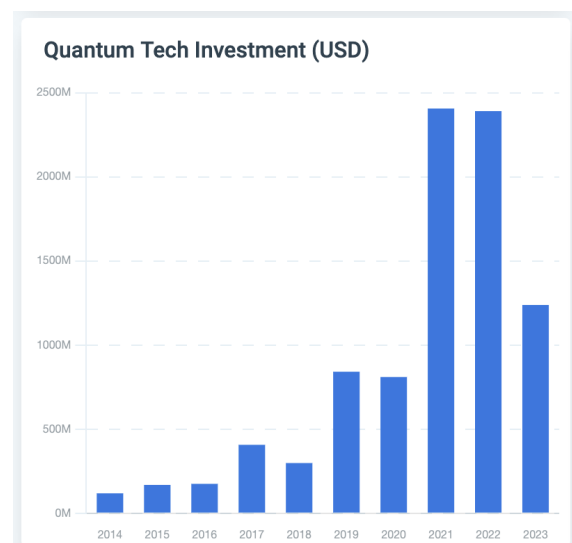


Figure 4 PhD student Alex Greene working on a superconducting quantum computing system at MIT (source: <https://news.mit.edu/2022/alex-greene-quantum-computers-1013>).

Last but not least, Quantum Technology is also a race for talents. To achieve this goal, 20 Universities from ten European countries are setting up an education and training program for quantum technologies in Europe with 16 new specialized Master's degrees: DigiQ (Digitally Enhanced Quantum Technology Master), the primary workforce development project of the Quantum Flagship (<https://qt.eu/>). But beyond training future quantum experts, Europe also needs to attract and keep the best talents of the world by offering attractive opportunities in its research and industry sectors.

Further reading:

- Is winter coming? Quantum computing's trajectory in the years ahead
- How Not To Invest Stupid And Other Smart Money Lessons From A Shark Tank Billionaire?
- <https://www.canva.com/design/DAFvjEJ4m8I/view>
- The Quantum Insider



Quantum explainer

Classical computers are built from classical bits, quantum computers use quantum bits, or qubits. And, of course, there is a fundamental difference between the two. A classical bit can have two values, which are usually designated as “one” (1) and “zero” (0). If we measure the state of a classical bit, it is either of these two values. If we repeat a calculation where (part of) the result is stored in a particular bit, then we find that the bit always has the same value (ignoring any errors in the classical computer).

But not so for quantum bits. A **quantum bit**, or **qubit**, has two fundamental states, which, again we designate with “one” and “zero”. But a qubit can also be in a **superposition** of these two states 1 and 0. And that is a fundamental difference between classical bits and qubits. However, when we observe a qubit, when we measure its state, the state of the qubit “collapses” to one of its fundamental states and we’ll see either a 1, or a 0. In other words, we cannot observe the **superposition** of the two states 1 and 0. And if we repeat the same quantum calculation, then we will find the qubit that stores the result sometimes in the 1 state and sometimes in the 0 state when we measure.

It is as if we have a box with a coin in it. Quantum mechanics tells us that the coin is showing both faces at the same time, each with a certain probability: it is in a **superposition** of heads and tails. But we cannot observe (measure) the coin without opening the box. But as soon as we do open the box, the state of the coin collapses to either heads or tails, not that superposition of states that it was in before we opened the box. Note that this is an explanation of superposition based on a macro scale system (a coin). Quantum mechanics deals with systems at the atomic level. The macroscopic explanation seems strange, or counterintuitive: how can a coin show heads and tails at the same time? In the macroscopic world that situation cannot arise, in the world of quantum mechanics, it can. The fact that quantum mechanics is counterintuitive makes it difficult to comprehend.

Quantum Gates: Classical bits are processed using classical gates. The simplest classical gate is the NOT gate, which negates the boolean value of its input: an input of 1 becomes an output of 0, and vice versa. Note that we can always deduce from the output of the NOT gate what its input was: the NOT gate is reversible. Two other types of classical gates are the AND gate and the OR gate. An AND gate produces a 1 on its output if all of its inputs are 1, otherwise the output is 0. An OR gate produces a 0 on its output when all of its inputs are 0, otherwise the output is 1. Note that the value of the inputs of the AND gate and the OR gate cannot be deduced individually from its output: classical AND and OR gates are not reversible, they “destroy” information. **Quantum gates** are always reversible. As their classical counterparts, they come in single qubit input form and in multi qubit input form. In general, single input quantum gates modify the superposition state of its input qubit. Multi input quantum gates take two or more qubits and modify their states based on their states. So, the superposition of states of the output qubits depends on the superposition of states in the input qubits: the states of the qubits on the output of a quantum gate are **entangled**.

Again, just like its classical counterpart, a quantum computer consists of sequences of combinations of quantum gates. At the moment most **quantum algorithms** are presented as such sequences of quantum gates, comparable, e.g., to a circuit diagram of an ALU (arithmetic and logic unit) in a classical computer.

Rydberg atoms are atoms of which one or more electrons in the outer shells are excited to higher energies. The higher energy states are closer together: the energy difference between these levels is much smaller than those of the lower energy states. That means that electrons in those higher energy states more easily travel between such states. Rydberg atoms can be arranged in arrays of interacting atoms. Through their natural interactions they can be used to simulate many-body quantum systems, a simulation that requires enormous classical compute resources.

Quantum supremacy: is defined as "the potential ability of quantum devices to solve problems that classical computers practically cannot" and was initially introduced by John Preskill. A paper published in Nature by a Google/NASA team [8], [9] claimed to have reached quantum supremacy. The fact is, if some sort of quantum advantage was demonstrated it was for a pretty useless application. If this work did not show that quantum computing is ready for real applications, it demonstrated that its basic principles are valid. Recent works shed another light on Quantum Supremacy by introducing theoretical elements showing that such a supremacy could be challenged on a classical laptop computer if limitations of real quantum hardware (noise, imperfections) are taken into account.

Hybrid Computing: Not all problems are fit for solving on a quantum computer. "Fit for" here means that the quantum computer can do exponentially better in terms of execution time than their classical counterpart. Sorting is an example: efficient classical algorithms for sorting are just as fast as their quantum counterparts. So, it is logical to combine classical and quantum algorithms in so-called hybrid algorithms to get the best of both worlds. An example of such a class are quantum variational algorithms. Such an algorithm requires a cost function which is used to find the global optimum in a multidimensional space. The computation of the cost function can in some of these problems more efficiently be computed by a quantum computer than by a classical computer, but the rest of the optimisation problem is more efficiently computed on a classical computer.

In the **Deutsch-Jozsa algorithm**, a computer must solve the following problem: find out whether a given function that maps multiple (n) boolean inputs to one boolean output is either constant, or is exactly balanced. A balanced function in this case means that it will produce true for exactly half of the number of possible inputs and false otherwise. And, of course, constant means that the function produces true for all inputs, or false for all inputs. The function is guaranteed to be balanced or con-

stant, but it is possibly rather computationally costly to evaluate. It can be shown that a classical computer needs to evaluate the function $2^n + 1$ times, while a quantum computer needs to evaluate the function only once. So, with increasing n , at some point the quantum computer will be able to solve this problem faster than a classical computer.

Further reading:

"Classical and Quantum Information Theory, An Introduction for the Telecom Scientist", Emmanuel Desurvire, Cambridge University Press, 2009, DOI: <https://doi.org/10.1017/CBO9780511803758> for a gentle introduction to quantum information technology

"IBM Quantum Learning", <https://learning.quantum-computing.ibm.com/>, (requires registration) for courses and hands-on experimenting with quantum computing.

Nielsen, M., & Chuang, I. (2010). *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511976667> for a thorough, theoretical introduction to quantum technology.

"The Feynman Lectures on Physics", Part III, Chapter 1-4 (https://www.feynmanlectures.caltech.edu/III_toc.html) for an introduction to the central ideas of quantum mechanics;

Quantum internet

A quantum internet is an application of quantum networks. Quantum networks enable the communication of qubits. Such networks can be used to connect quantum processors to form more powerful quantum computers. Quantum networks can also be used to create quantum internet applications. One such application is the secure distribution of cryptographic keys: in this setup, cryptographic keys are distributed over a quantum network using entangled qubits, enabling the detection of eavesdropping on the communication. But quantum internet, just like quantum computers, are under development and are still much in the research phase. Practical applications at this moment require complicated setups, often involving cryogenically-cooled devices, preventing widespread use today and in the next few years. (Singh, Dev, Siljak, Joshi, & Magarini, 2021)

Conclusion

Quantum computing will have an indirect impact in the next few years with respect to embedded software development and embedded software. It depends on the speed of evolution and innovation of quantum technology when quantum computing devices will leave the laboratory and make their introduction to the industry. For now, that appears to be at least a decade away, but vigilance on this subject is required. And Europe should strive for independence from other nations in this area to be able to develop this technology on its own, in the light of the recent developments in international relations.

References

- [1] R. P. Feynman, "Simulating physics with computers,," *Int. J. Theor. Physics*, pp. 21, 467-488, 1982.
- [2] J. Preskill, "Quantum Computing in the NISQ era and beyond,," *Quantum*, p. 79, 2018.
- [3] "NUKIB - Tsjech National Cyber and Information Security Agency," [Online]. Available: <https://nukib.cz/cs/infoservis/aktuality/1984-nukib-pripravil-podpurne-materialy-pro-ochranu-pred-hrozbou-v-podobu-kvantovych-pocitacu/>.
- [4] T. Attema, J. Diogo Duarte, V. Dunning, M. Lequesne, W. van der Schoot and M. Stevens, "Het PQC-migratie handboek (in Dutch)," 2023.
- [5] "Post-Quantum Cryptography - Setting the Future Security Standards," [Online]. Available: <https://www.nxp.com/applications/enabling-technologies/security/post-quantum-cryptography:POST-QUANTUM-CRYPTOGRAPHY>.
- [6] A. Singh, K. Dev, H. Siljak, H. Joshi and M. Magarini, "Quantum Internet—Applications, Functionalities, Enabling Technologies, Challenges, and Research Directions," *IEEE Communications Surveys & Tutorials*, pp. 2218-2247, 2021.
- [7] B. Kantsepolsky, I. Aviv, R. Weitzfeld and E. Bordo, "Exploring Quantum Sensing Potential for Systems Applications," *IEEE Access*, pp. 31569-31582, 2023.
- [9] <https://www.nasa.gov/technology/computing/google-and-nasa-achieve-quantum-supremacy/>
- [9] Arute, F., Arya, K., Babbush, R. *et al.* Quantum supremacy using a programmable superconducting processor. *Nature* **574**, 505–510 (2019). <https://doi.org/10.1038/s41586-019-1666-5>
- [10] <https://quantera.eu/quantum-technologies-public-policies-2023/>
- [11] D. Vion et al., "Manipulating the Quantum State of an Electrical Circuit," *Science*, vol. 296, no. 5569, pp. 886–889, May 2002.
- [12] J. I. Cirac and P. Zoller, "Quantum Computations with Cold Trapped Ions," *Phys. Rev. Lett.*, vol. 74, no. 20, pp. 4091–4094, May 1995.

Artur Garcia-Saez is a researcher at the Barcelona Supercomputing Center (Spain), and a co-founder of Qilimanjaro Quantum Tech.

Christian Gamrat is a researcher in the Research and Technology Department at CEA (Alternative energies and Atomic Energy Commission), France.

Harm Munk is project leader at TNO, The Netherlands.

Paul Carpenter is a researcher in the Computer Sciences Department at Barcelona Supercomputing Center, Spain.



Open source drives collaboration and innovation. It is well established for software but still in its early days for hardware. Open source hardware will lower the barrier to entry and help achieve European sovereignty and reduce vendor lock-in.

Open hardware and RISC-V: an exciting opportunity for Europe

by Miquel Moretó, Osman Unsal, Adrian Cristal, Jérôme Quévremont and Gaël Blondelle

With economies discovering the limits of the globalisation of the past three decades, the world is entering a new phase. In Europe, this phase is further driven by the need to protect European citizen's security, and to assert sovereignty in key strategic technologies. Particular key technologies are semiconductors and computing. To retain the capability to compete in this key technology, European nations and companies have to rely on a set of common building blocks. These building blocks are best served through creating an open hardware environment.

Europe has traditionally been strong in building industries that are greater than the sum of their parts. This collective ecosystem is driven by small-to medium-sized enterprises (SMEs) as much as by larger companies. In this sense, an open hardware environment is the best avenue to enable these myriad actors to march in unison towards the main goal of creating a world-class hardware base.

The focus of this article is open source IP and chip design, including all the layers of the hardware compute stack from high-level synthesis languages down to EDA tools development. There are multiple initiatives for open hardware, examples include RISC-V ecosystem (<https://riscv.org>), Open Compute <https://www.opencompute.org/>, Open Hardware Repository (<https://ohwr.org>) and Open Source Hardware Organization <https://www.oshwa.org/>.

Among the alternatives, the RISC-V path has gained the most traction in Europe and accordingly this report proposes RISC-V as Europe's open hardware solution of choice.

Key insights

- Open source is a global framework for collaboration and innovation. It is well established for software, but still at an early stage for hardware. Open source delivers benefits across the whole landscape from HPC to cloud to IoT and AI.
- Open source hardware comprises open hardware interfaces including the Instruction Set Architecture (ISA), the high-level hardware descriptions of core components (e.g. processing cores, caches, and peripheral memory controllers), process for fabrication, and electronic-design automation tools.
- It is insufficient to merely release the hardware, as abandonware, in an open source repository to satisfy the funding agency. Open source foundations like the Eclipse Foundation and OpenHW Group are there to support research and industry players to create a real open source community and visibility.
- Hardware has higher requirements for verification than software, with peculiarities to ad-

dress by the governance of open source hardware projects. Hardware verification is more costly than development, so once verification is done, the design will likely be frozen, and should not evolve.

- Open source collaboration enables a high degree of software/hardware co-design due to open interfaces and implementations across the whole software/hardware stack. An example is the collaboration between OpenHW and RedHat [5].
- Open source hardware allows accessing source code (e.g. for certification) achieving such sovereignty, avoiding export controls and without depending on a specific vendor. Sovereignty in hardware design is especially strategic for cybersecurity, safety critical (e.g. automotive), defense and other specific sectors affecting national security [6].
- Open source lowers the barrier to entry. In particular, permissive open source licensing (Apache, MIT, BSD...), as selected by the OpenHW Group, allows the integration of open source components in proprietary designs. In contrast, a GPL licence on IP forces the whole design to become GPL, since hardware has no equivalent to software's dynamic loading. This leads to dual licensing approaches where the design made available under GPL is fully owned by a company that licences it under proprietary licensing terms for commercial use. The dual licensing model can enable the freedom for academic use when a company wants to be able to capture some of the value from commercialization of successful designs.

Permissive licensing recently proved successful in the open source hardware realm, e.g. through the adoption of the academic PULP cores by OpenHW Group projects to mature them to production-grade IPs. This permissive open source licensing approach is gradually replacing the dual licence model.

- Among open source hardware, RISC-V has so far gained the most traction in Europe.
- RISC-V and open source hardware are great tools for use in education and academia, to help attract and retain talent.

Key recommendations

- Europe needs to establish truly European solutions for computing. Open source hardware and RISC-V are seen as good leverages to increase European sovereignty and economy, in conjunction with the European Chips Act.
- We propose RISC-V as Europe's open hardware solution of choice for all domains ranging from embedded systems to HPC.
- Hardware developers should release as much IP as possible in open source in order to promote collaboration between academia and industry.
- We recommend the adoption of permissive open source licences, which have been more successful in driving adoption, and lowering barriers to collaboration between research, academia and industry.
- Developers should leverage the experience, governance, and processes of existing open source foundations established in Europe in order to build an open source community and visibility.
- Hardware engineers and students should be aware of open hardware and should be encouraged to contribute to open source hardware during their career or studies, as is common for software.

Open source governance and community building

The term "Open Source" was defined in the late 90's by the Open Source Initiative (OSI), a created by a group of entrepreneurs that wanted to promote Free Software without using the term free, which is ambiguous in English with meanings both about "freedom" and "gratis". An open source license is one that has been accepted by the OSI as being compliant with the open source definition, which emphasises no discrimination against persons or groups, and no discrimination against fields of endeavour, source code access, freedom to create derived works, and free redistribution.

Over the last 25 years, open source has enabled rapid innovation in the software industry and has become the preferred way of setting up global collaborations and ecosystems for core technologies and non-differentiating software. An enabler of

this global collaboration is that open source licences, like the Apache Software Licence, the MIT or BSD licence for example are short and easily understood and applied worldwide. Those “permissive licences” have proven to be efficient to enable collaboration between research, industry and startups.

The bloom of open source has been supported by the rise of open source foundations as they developed their well structured open source governance framework as a set of rules and practices that guide the development of open source projects and the management of open source communities. For example, the Eclipse governance model is supported by an open and transparent development process that supports decision making, ensures vendor neutrality and provides world-class intellectual property management.

Strong open source governance supports growing a community because the rules of engagement are clear. Over the decade, most corporations have defined policies to allow their employees to contribute to open source software and it is time to extend this to open source hardware. As open source hardware is recent, most hardware developer teams have no prior open source experience and face a longer learning curve.

In the last decade, the principles of OSS governance have been extended to open source hardware, particularly in computing architecture. Indeed, much of modern hardware development, especially in areas like RISC-V processors, resembles software development. Thus, groups like the OpenHW Group are adopting open source models for hardware development, using licences like Apache/Solderpad.

But open source hardware projects are different from open source software, for example due to specific verification requirements, which has an effect on the way that the hardware description codebase is managed. The OpenHW Group’s industrial projects prepare intellectual property (IP) blocks (RISC-V cores, accelerators...) that will eventually get into integrated circuit designs. Once etched in silicon, any bug can cause catastrophic consequences. It is therefore necessary to fully verify open source hardware IP blocks before releasing them. Verifying a RISC-V processor core represents significantly more effort than designing it. Therefore, once the core is fully verified, the source code will likely be frozen, and should not evolve for that design.

The rise of an open Instruction Set Architecture

The advent of Artificial Intelligence applications in the embedded system market has raised the demand for high performance computing capabilities in embedded processors. This demand is presently coming from cloud based computing services for embedded applications and will involve more and more into edge computing devices featuring high computing power close to the final user or even to Internet-of-Things nodes. This reality opens a particularly favourable scenario for horizontally sharing the same computing platforms (instruction set architectures) in different contexts, ranging from the embedded market to the HPC market.

In this view, the appearance of the RISC-V instruction set (Patterson, 2016) (Waterman, 2013) on the embedded system scene is of particular interest, as it allows processor designers to join the high volumes of embedded solutions with the advantages of an open instruction set and software technology. As the software stack support matures for different market segments, RISC-V-based processors are making the jump from embedded systems to data-centre class CPUs and Supercomputing.

RISC-V originated in 2010 from a research project at the University of California, Berkeley and is now supported by the RISC-V International foundation counting about 4,000 members, among which are numerous major industrial actors in the ICT market (<http://riscv.org>). RISC-V is composed of a base instruction set—divided into user and privilege sets—that has been finalised and will never change, extended in a modular fashion by a number of dedicated instruction sets targeting higher performance or specialised application domains. Interestingly, there is a vector processing extension, already ratified with the release of RVV1.0 (<https://wiki.riscv.org/display/HOME/Recently+Ratified+Extensions>). The latest version of the extension is optimised for embedded systems, and ongoing discussion is aimed at optimising vector support for other domains such as HPC and AI. The matrix multiplication extension, expected to be defined and ratified in 2024; will expand the RISC-V applicability towards hardware support for modern AI applications. This illustrates the positive impact of open specifications evolving to address the needs of large communities.

The RISC-V initiative inherits the long history of RISC processors started in the 1980s from the research initiatives led by John Hennessy and David Patterson. In 2018, they received the ACM Turing award “for pioneering a systematic, quantitative approach to the design and evaluation of computer architectures with enduring impact on the microprocessor industry”. It is important to note that while RISC-V is an open ISA standard, this does not automatically imply open source hardware. However RISC-V has been the ISA of choice for most open hardware efforts, since the RISC-V ISA standard is free for both senses of the word.

At present, while several RISC-V proprietary and open products are emerging for the embedded market, there is limited availability of RISC-V HPC processors, either implementing vector computation or targeting general purpose HPC applications. The 45 nm vector-accelerated University of California, Berkeley Hwacha chip achieves 16.7 Double-Precision GFLOPS/W, while the 28 nm FDSOI Hwacha chip employs aggressive voltage biasing techniques to achieve power efficiency in variable load conditions. The SiFive’s U540 Quad-Core Processor addresses general purpose performance and exhibits a 6 GFLOPS theoretical peak at a 4.63 W board power consumption, showing the need for accelerator support to target breakthrough power efficiency.

Sovereignty

In the past years, a few events have revealed European weaknesses and the need to regain control of key electronic sectors. The 2020 semiconductor shortage, which followed the COVID lockdowns, severely affected the European industry, in particular critical sectors such as automotive. The projected acquisition of Arm by NVIDIA in 2021 demonstrated to semiconductor manufacturers that a competitor could take the control of a then quasi-monopolistic supplier for their embedded computing architectures.

“Sovereignty” can have several embodiments: gaining full control over a critical asset, ensuring a sufficient availability of a critical asset, or making sure that no-one can prevent you from exploiting a critical asset. In the case of semiconductor manufacturing, the European investments deriving from the Chips Act are a way to increase control on the semiconductor supply chain for the sake of

the European economy. In the case of open source hardware for computing solutions, the goal is to support the free flow of the technology, and ensure that European actors cannot be blocked from the use of a critical intellectual property.

A representative illustration of the protection against export restrictions that can be brought by open source lies in the ongoing economic confrontation between the USA and China. In 2019, the Chinese manufacturer Huawei was prevented from integrating Google proprietary apps (Maps, YouTube...) into their smartphones by then-president Donald Trump [1]. But he could not prevent the Chinese company from integrating Android, as the latter is available as open source. The Linux Foundation published a report [2] to explain and elaborate on open source and US export control [4].

On our continent, the European Commission has mandated a working group on sovereignty and published their “Recommendations and Roadmap for European Sovereignty in Open Source Hardware, Software, and RISC-V Technologies” [3]. From the conclusions, the Key Digital Technologies Joint Undertaking (KDT JU, now Chips JU) has carried out several calls for projects, which gave birth to the TRISTAN and ISOLDE open source hardware projects. Other calls and projects are expected to follow to increase the European impact in the open source hardware field.

Beyond the protection against quasi-monopolies and geopolitical issues, open source unleashes new avenues for business and innovation, through participation and collaboration of many stakeholders working towards the same goal – European leadership. Finally, adopting and contributing to open source hardware creates technical and business differentiators: the ability to customise and get a better architectural/energy/performance fit for one’s application; ability to perform white-box analyses for critical domains; sharing costs instead of purchasing IP... A stronger economy is a good factor to defend the future of our continent.

Finally, it is interesting to note that RISC-V International, the foundation that maintains and develops the ISA, has moved to Switzerland, which grants neutrality from national regulations.

Conclusion

This position paper has defined open source hardware, advocated for an open source hardware ecosystem based on the RISC-V standard, surveyed the state of the art in RISC-V open source hardware; and argued that RISC-V is a good choice to ensure European sovereignty.

References

- [1] BBC, "Huawei's use of Android restricted by Google", 2019 [Online]. Available: <https://www.bbc.co.uk/news/business-48330310>. [Accessed 23 November 2023].
- [2] The Linux Foundation, "Understanding Open Source Technology & US Export Controls", 2021 [Online]. Available: <https://www.linuxfoundation.org/resources/publications/understanding-us-export-controls-with-open-source-projects>. [Accessed 23 November 2023].
- [3] Open Source Hardware & Software Working Group, "Recommendations and Roadmap for European Sovereignty in Open Source Hardware, Software, and RISC-V Technologies", 2022 [Online]. Available: <https://digital-strategy.ec.europa.eu/en/library/recommendations-and-roadmap-european-sovereignty-open-source-hardware-software-and-risc-v>. [Accessed 23 November 2023].
- [4] Reuters, "RISC-V technology emerges as battleground in US-China tech war", October 2023 [Online]. Available: <https://www.reuters.com/technology/us-china-tech-war-risc-v-chip-technology-emerges-new-battleground-2023-10-06/> [Accessed 23 November 2023].
- [5] OpenHW, "OpenHW Group Announces CORE-V CVA6 Platform Project for RISC-V Software Development & Testing", November 2023 [Online]. Available: <https://www.openhwgroup.org/news/2023/11/07/openhw-group-announces-core-v-cva6-platform-project-for-risc-v-software-development-and-testing/> [Accessed 23 November 2023].
- [6] Open Source Hardware & Software Working Group, "Recommendations and roadmap for European sovereignty on open source hardware, software and RISC-V Technologies", September 2023 [Online]. Available: <https://digital-strategy.ec.europa.eu/en/library/recommendations-and-roadmap-european-sovereignty-open-source-hardware-software-and-risc-v> [Accessed 23 November 2023].

Miquel Moretó leads the High Performance Domain Specific Architectures team at Barcelona Supercomputing Center.

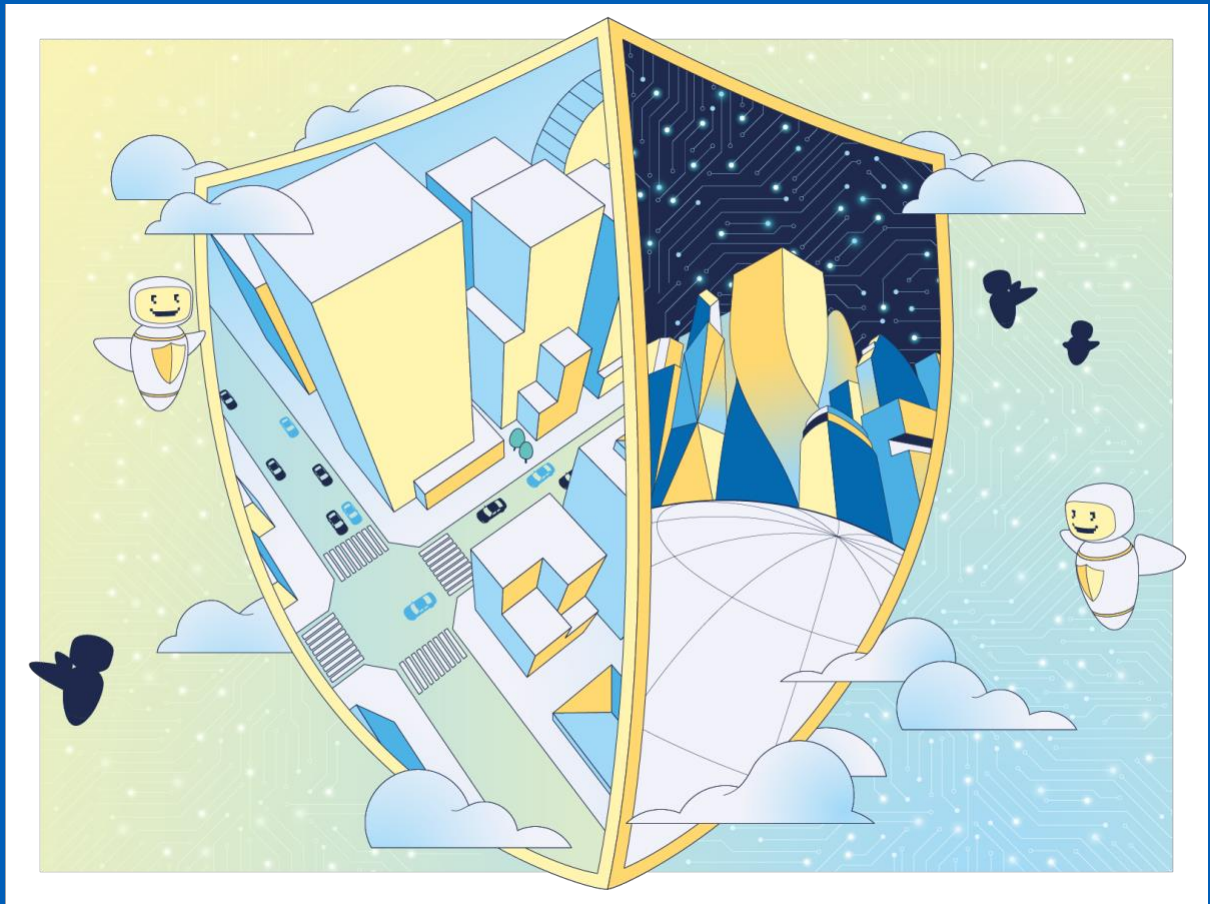
Osman Unsal is co-manager of the Computer Architecture for Parallel Paradigms research group at Barcelona Supercomputing Center. His group has developed in-order and out-of-order RISC-V cores; as well as the RISC-V vector processing unit in the European Processor Initiative project.

Adrian Cristal is co-manager of the Computer Architecture for Parallel Paradigms research group at Barcelona Supercomputing Center.

Jérôme Quévremont is RISC-V and open hardware project leader at Thales Research & Technology. He serves as co-chair of the Technical Working Group at the OpenHW Group and the chair of the Functional Safety special interest group at RISC-V International.

Gael Blondelle is chief membership officer at the Eclipse Foundation AISBL. He has been working in open source since 2004 and is helping organisations collaborate in open source since he joined the Eclipse Foundation in 2013.

Cybersecurity





The Next computing Paradigm (NCP), with its massively interconnected components, faces compounded cybersecurity and privacy issues, at all levels of its continuum. These must be addressed for the concept to succeed.

The race for NCP cybersecurity

by Olivier Zendra and Bart Coppens

The Next Computing Paradigm (NCP), characterized by its massive interconnectivity between numerous services and systems, faces significant cybersecurity and privacy challenges across its entire spectrum.

The vast number of interconnected systems and services offer a very large attack surface. Given the pervasiveness of the NCP, reaching the cyber-physical world, the potential consequences of attacks include not only the risk of leaking sensitive data and potentially leading to large economic losses, but attackers could also gain the capability to yield tangible and, in some cases, life-threatening impacts.

Implementing the NCP thus extends way beyond facilitating harmonious interactions among its myriad services, which is a challenge in itself. It fundamentally hinges on ensuring the cybersecurity of this intricate network. The acceptance of the NCP is contingent on trust, and to garner trust, effective security must be guaranteed. Significant challenges persist in establishing the necessary levels of security for the NCP, and these challenges are of critical importance to pave the way for the paradigm's acceptance and success.

This chapter contains six contributions related to cybersecurity for the NCP, four of them more related to software security and two to hardware security:

- **“The NCP cybersecurity challenges”.** This article provides an overview of some of the cybersecurity challenges the NCP faces,

including classical software supply chain issues, as well as challenges in this context of massively interconnected elements. The case is also made for the vulnerability of two specific domains of interconnected elements belonging to the NCP, namely critical supply chains and services (like gas and electricity supply) and healthcare infrastructures and hospitals. If the EU wants to be strong, it needs strong cybersecurity on these aspects, now.

- **“More data for the NCP implies more privacy risks”.** This article addresses the privacy issues induced by the multitude of data and communication in the NCP. Indeed, the latter includes numerous home sensors that will communicate with servers and services, and that could very easily leak private information if not properly taken care of, impairing the NCP acceptance by people.
- **“The browser: the key to your privacy on the Web”.** This article addresses browser tracking issues. Currently, and in the Next Computing Paradigm (NCP) as well, one of the major ways of interacting with the Web is through the browser. But tracking users through their browser is commonplace, and often underestimated, with the user not being aware of them. The browser is thus a keystone to user privacy in the NCP.
- **“DLT and IPFS Technologies are Paving the Way for the NCP”.** This article explains how distributed ledgers and file systems can provide the trustability and resilience to the NCP through decentralization, immutability,

and transparency of interactions and transactions, removing single points of failures. However, significant challenges remain, such as scalability, interoperability, and user adoption.

- **“Integrity at Every Link: A Roadmap to Trustworthy Hardware Supply Chains”.** This article makes the case that hardware constitutes the foundation of any computer system, and thus its integrity is crucial for cybersecurity. However, ensuring this integrity throughout the entirety of the hardware supply chain poses a significant challenge in establishing a secure computer system.
- **“Microarchitectures as Root-of-Trust in Computing Systems – Research Needs in Formal Security Analysis”.** This article explores using formal methods for verifying hardware security, a shift from the traditional software-centric approach. While computing systems rely on hardware as a "root-of-trust" for security, new vulnerabilities exist that cannot be patched in software alone. Thus, security verification and the development of defense mechanisms must be pursued at the microarchitectural hardware level, utilizing formal methods specifically tailored for security verification.

Key insights

- The NCP comprises many small components that are discovered and integrated at run time. This leads to cybersecurity risks, in software supply-chain and in inter-components interactions.
- A tremendous number of IoT devices are unsecure, especially communication-wise.
- LLM security is in its infancy.
- Many innocuous data sources like IoT-based sensors can be used alone or in combination to infer personal and private information. The NCP will merge multiple data sources, which may present privacy risks if not properly designed and implemented.
- User tracking through the browser is extremely widespread. Users are tracked almost all the time, on most websites. But browser fingerprinting has security uses that websites actively use.
- Vendors add new functionalities to browsers at a fast pace, but not enough effort is put into minimizing privacy risks.
- DLTs and IPFS are key enablers to an overall decentralized secured platform on the Internet.
- Mass adoption of this new computing paradigm will require thorough transformation of European laws and policies, including on the legal value of smart contracts.
- Design-dependent hardware Trojans are a fundamental security issue. Standard detection systems only identify known hardware Trojans.
- Hardware microarchitecture security flaws expose a very large attack surface. Vulnerabilities often arise from integrating multiple components and specific hardware-software interactions.
- Ad hoc processes that require design changes and software developer collaboration are used to mitigate hardware security issues. Security methods to address timing side channels increase manual design workload and hardware overhead.

Key recommendations

- Promote privacy-enhancing technology research to prevent data breaches or lower their impact.
- To enhance privacy, systems should be designed to not leak any data without explicit and meaningful user consent. No backdoor should be allowed.
- Invest in methods and tools to have security/privacy as a first-class citizen during the development of NCP hardware and software elements, including quantitative security metrics/properties.
- Invest in methods and tools to find vulnerabilities in existing IT systems (e.g., with static analyses on source code and behavioural analyses at runtime), and automatically prevent or mitigate them (e.g., with automated refactoring tools and blocking systems),
- Invest in automated methods and tools, possibly based on AI, for runtime detection of intrusions, attacks and privacy breaches, and active cyber defense.

- Base the critical parts of the NCP on understood open-source software and hardware, or on EU-made, trustable because audited software and hardware.
- Invest in authentication/identification of people and NCP software and hardware elements, both for end-to-end trust and tracing of suppliers
- Evaluate and improve blockchains/DLTs security and performance/scalability, both at algorithmic level and implementation level.
- Investigate the security of LLMs, with a particular focus on LLM prompt security.
- Mandate securing the radio communications of IoT systems.
- Expand mandatory EU-based, multi-level NCP elements security certification, holding IT systems providers and resellers liable for poor security, while taking the specifics of open-source projects into account.
- Invest in methods and tools that break the uniqueness of people when browsing and allow users to control fingerprinting techniques.
- Encourage browsers to block third-party cookies.
- Enforce GDPR wrt. user tracking, ensuring collected data is strictly necessary to the usages accepted by users.
- DLTs and IPFS allow combining security and controlled openness. EU companies and public authorities (including regulators) should thus rethink their models, moving from “paper-based policies” to “code-base policies” relying more on mission-oriented consortiums than competition.
- Investigate and monitor possible ways, especially security issues, to circumvent European policies using DLTs and IPFS technologies, whether by European or foreign actors
- Invest in long-term research to secure the hardware supply chain and in short-term, best-effort security to protect against malicious hardware Trojans:
 - automatic, zero-fault, non-destructive reverse engineering methods enabling end-to-end equivalency validation, and complete equivalence checking methods from design specifications to physical devices.
 - active measures to prevent malicious design modifications in hardware supply chains.
- Invest in methods and tools for SoC microarchitecture security analysis, formalising micro-architectural threat models for new non-functional formal verification approaches.
- Invest in verification-driven, secure-by-construction design providing system-wide threat coverage.
- Promote open-source and public-domain initiatives, such as the RISC-V ecosystem.

Olivier Zendra is a Tenured Computer Science Researcher at Inria, Rennes, France.

Bart Coppens is a part-time assistant professor and a post-doctoral researcher in the Electronics department of Ghent University, Ghent, Belgium.



The Next Computing Paradigm (NCP), with its numerous, interconnected and communicating elements, from servers and services to IoT, sensors and user interfaces, offers an immensely vast cyberattack surface. This must be addressed for the NCP to succeed.

The NCP cybersecurity challenges

by Olivier Zendra and Bart Coppens

Key insights

- NCP services are vulnerable to software supply-chain attacks, especially with the multiplication of components providers.
- The very large number of services composing the NCP creates numerous integration and communication points between these services that have not been evaluated together for security, hence potential vulnerabilities.
- The *dynamic* composition of services creates further attack opportunities.
- The NCP reaches the physical world, with significant and impactful cybersecurity issues in critical supply chains and services, like gas and electricity supply and healthcare infrastructures and hospitals.
- A tremendous number of IoT devices are unsecure, especially communication-wise.
- LLMs security is still in its infancy.

Key recommendations

- Promote research and tools for finding and removing vulnerabilities in software supply chains.
- Promote research into the cybersecurity of systems with many small components that require secure interactions, especially in cases when the interacting components are unknown.
- Promote the cybersecurity of critical physical supply chains and services, like gas and electricity supply, and healthcare infrastructures and hospitals.
- Promote IoT cybersecurity, especially with regards to radio communications.
- Promote research in LLM security, especially and urgently against prompt poisoning attacks and leakage of information.

The NCP, an extended playground for cybervillains

Digital technology is increasingly altering the interactions of EU citizens with governments and corporations. The reliance of EU citizens on the current computing continuum, and on the coming NCP, including smartphones, smart devices, large computer systems, cloud, edge, and IoT devices, is ever growing. Social media and electronic commerce streamline and individualise customer service. Today's EU citizen disregards borders when purchasing goods, services, and commodities online. Similarly, contemporary domestic appliances have been digitized and computerised. Internet-controlled HVAC systems are a feature of smart residences, and 5G/6G smart networks will accelerate this development. The Internet of people and the Internet of machines (Industry 4.0) interweave. The risks associated with the digitalization of daily life include the potential for private data exposure or harm to critical infrastructure due to security weaknesses in software systems. Such incidents have the capacity to endanger life or state sovereignty. Consequently, cybersecurity is a crucial aspect, all the more in an NCP composed of many interacting devices and services, coming from many suppliers.

In recent decades, numerous tools and solutions have been developed with the intention of finding, reducing, or eliminating vulnerabilities such as the ones that can be found in the NCP. Static analysis [15,16], fuzzy testing [17,18], compiler-based mitigations [19], verified compilation [20], symbolic execution [21] and other such techniques are among these. However, due to the difficulty and expense of recruiting a highly qualified and specialised workforce, these techniques are seldom implemented and are sometimes unattainable. Moreover, the intricacy of complex systems, such as the ones that compose the NCP, frequently renders these approaches unworkable, even when employing developers with exceptional expertise. Decades-long, unsecured code bases continue to be a significant concern throughout the NCP computing continuum.

NCP and software supply chains attacks

In an NCP composed of many interacting services, coming from many suppliers, and even more numerous software repositories and supply

chains, the sheer amount of source code for components, or even binary libraries, represents a tremendous attack surface for villains.

Indeed, security vulnerabilities exist in nearly all computing software, and attackers exploit these to circumvent the host's security, which can result in dramatic consequences. Software supply chains, which include software and its dependencies, both for production and distribution, are also subject to cybersecurity concerns. It has been demonstrated that contemporary software development, which encourages the reuse of (open-source) components, fosters innovation. This is completely logical: in this manner, software developers can specialise in their own domain while simultaneously capitalising on the expertise of others. This approach streamlines and reduces the cost of software development by utilising pre-existing code bases for numerous functionalities. Many modern software projects thus do not create every line of code from scratch, but rather depend on a multitude of other projects, libraries and frameworks. Additionally, this could enhance security, e.g. by eliminating the need for an app developer to have a Ph.D in cryptography and related disciplines to establish a secure network connection for their application. However, new cybersecurity concerns have emerged due to the increased diversity and complexity of the software, its production and deployment. This is due to the fact that security weaknesses in an extensively utilised library may impact all systems that employ this library. In fact, *software supply chain* attacks further compromise security by injecting or committing malicious code into software repositories and build chains, which then unknowingly distributes the malicious code via the standard development and distribution processes.

The Log4j vulnerability [1] is one example of such a software supply chain vulnerability. Revealed in December 2021, it could provide attackers full control of a system and sent cybersecurity stakeholders rushing [2]. Log4j is a popular logging package for Java applications that is utilised in many products as a common off-the shelf tool that many use without paying too much attention to it. The potential impact was significant yet initially unknown. Mitigating factors prevented a worst-case scenario [1], such as the fact that not all versions were affected, only certain configurations allowed tampering with the affected versions, and some systems were using a non-vulnerable Log4j API despite integrating it. This can be regarded as a free warning of what could have happened when

an apparently unimportant, known, but widespread component is vulnerable.

The “2020 United States federal government data breach” [3] with the SolarWinds attack [4] is another famous example. This included a supply chain attack on Microsoft cloud services and another one on SolarWinds' widely used Orion software. Other product vulnerabilities gave attackers more access. This cyber-espionage event is the worst in U.S. history because of its duration (8-9 months) and the targets' sensitivity and relevance. At least 200 businesses may have lost data due to the attack. NATO, the UK and the US governments, European Parliament, Microsoft, and others [3] were affected.

Phishing attempts have also targeted Python Package Index contributors, a big software repository utilised by developers worldwide. Then attackers could use the phished credentials to upload malicious versions of the packages maintained by the phished users [5].

External attackers are the most typical and acknowledged source of supply chain threats. However a legitimate developer can also go rogue and inject malicious changes. A JavaScript library developer is an interesting recent example: when the war between Russia and Ukraine began, this developer changed their code to identify if it was placed on a computer with an IP address geolocated in Russia or Belarus and, if so, wipe files [11]...

As can be seen, attacks through the software supply chain are varied and common. The combination of many services in the NCP makes the magnitude of secured components even greater. Without proper expertise, experience and tools, it could become exceedingly challenging, if not unattainable, for software developers to create secure applications. Thus, providing tools able to automatically find, mitigate, and correct security vulnerabilities in source code, in a scalable way, is crucial to the NCP security.

Integration and Communication of services in the NCP

The NCP is a system composed of numerous components that must communicate with one another. Due to the dynamic and modular nature of the NCP, it will be impossible to determine at the time of design or implementation which specific

services will interact with one another in a given system. This has some security-related repercussions.

First, the inclusion of numerous small components in a system may potentially enhance its security by facilitating a distinct separation of concerns. Furthermore, in adherence to the principle of least privilege [6], only those privileges necessary for each component to perform its task can be granted to it. This may potentially mitigate the effects of an attack by restricting the attacker's access to that component's capabilities.

Splitting a system into numerous components is not, however, a security panacea. Privileges that an attacker desires (access to sensitive information that could be leaked, access to a cyber-physical component that could injure people, etc.), despite being restricted to specific components, will continue to exist throughout the entire system. This first implies that it continues to be critical that every component, privileged or not, be designed and developed in a secure manner, making it difficult for an attacker to compromise a component. However, regardless of how circumscribed or limited its privileges may be, a component must still engage in communication with other components. An attacker may still attempt to leverage this. Thus developers must bear in mind that interactions with other components are not reliable, as they may be susceptible to compromise by an attacker.

Unfortunately, actual attacks do occur against systems that have been designed to be secure by being divided into multiple components. Such an attack can occur when a portion of the codebase lacks trust in the operating system but still requires security guarantees. This means that a segment of the code is isolated from the operating system, that it does not trust. However, this code still needs to interact with the outside world, and requires the operating system to mediate this. The untrusted operating system thus interfaces with the protected enclave so as to provide inputs to it, thus undermining its security goals [7,8]. Similarly, when compartmentalizing a single application into smaller components with limited privileges, it does not suffice to have one component per library, with the original library's interface working as the interface between compartments. Indeed, these interfaces are typically not designed as explicit trust boundaries [9]. An attacker could thus exploit one library/compartment, and use that to force other compartments/libraries to leak data or corrupt data, which could lead to total attacker

control, loss of confidential information, or disruption of service on the targeted library/component [9].

To evaluate the security of such a system, a first step will be to evaluate the security of the individual components. One advantage of having small components is that it should be easier to evaluate the security of each of them, rather than that of a large monolithic system. But although each component is small, the system as a whole is now composed of a large number of such components, necessitating numerous evaluations. Moreover, the absence of co-development among the various components may complicate the evaluation of the *combined* system's security, since the full scope of the system is not known at development time. Thus, it is important that research continues into analysing and improving the security of systems that consist of many small components that need to securely interact with one another, especially when it is unknown upfront exactly which components will interact.

The NCP and physical world cyberattacks

The NCP is not only all the code and software components that form it. The NCP does extend its reach to the physical world, through CPS (cyber physical systems), that are computing systems intertwined with physical systems. As such, attacks against the NCP do reach or specifically the physical world too, especially physical critical infrastructures.

Physical supply chain attacks have become (in)famous. The Aurora Generator Test [10] by the Idaho National Laboratory in 2007 first publicly demonstrated how a cyberattack could damage a 2.5MW diesel generator on the electric grid. But the first publicly widely recognised successful cyberattack on a power grid outside of a lab was the 23 December 2015 hack on Ukraine's power infrastructure [12], which left 230,000 people without power for several hours. Critical supply chain cyberattacks like this stand out. This sophisticated operation was planned and executed for months, meticulously gaining access and placing triggers in the Ukrainian electrical grid. It used many techniques, including trojans, viruses, and 1990s Microsoft Word macro-based malware. The latter proves that even dated attack vectors may cause harm. It was a full attack on both customer-delivery systems and distribution centre backup

systems to blind power technicians. It also included a DDoS attack on the operators' call centres to prevent customers from reporting the problem's breadth and getting informed. Several factors suggested a Russian strike, while attribution is still unclear [12].

The “NotPetya” malware is considered the “most devastating cyberattack in history” [14]. It targeted “complete energy companies, the power grid, bus stations, petrol stations, the airport and banks” [13], critical infrastructures and mostly supply chains. World leaders including Merck, TNT Express (European division of FedEx), Maersk, DHL, India's largest container port JNPT, food firms, and others are among them. Restoration of minimum operations took days and full operations months for affected companies.

In addition to these very widespread attacks in physical supply chains, scores of more localized, yet devastating attacks have taken place in other vulnerable areas inside the NCP, especially on the *healthcare infrastructures and hospitals*.

Much evidence exists of major attacks on healthcare infrastructures throughout Europe [23]. For example, the National Health Service (NHS) of the United Kingdom was significantly impacted by the 'WannaCry' ransomware attack in 2017. The Health Service Executive (HSE) of Ireland was compromised in May 2021 by the 'Conti' ransomware. Compared to 2019, the number of successful cyber attacks targeting critical infrastructure health service providers in Germany more than doubled in 2020. In 2020, 27 significant cyberattacks were directed at French health institutions. Throughout the entire year of 2020, Spain's health sector was subject to frequent attacks, with reports of as many as 50,000 harmful incidents, of which 375 were successful [23].

The following are some tangible, specific incidents that have occurred in recent years (this list is in no way exhaustive). The hospitals in Villefranche-sur-Saône and Dax, France, were targeted by ransomware attacks in February 2021 [29]. Critical departments, including radiotherapy, radiology, the laboratory, pharmacy, automated washing cycles, and room catering, were significantly impacted by these incidents. As a consequence, operational activities were significantly disrupted, leading to treatment delays and patient rerouting. Surgical procedures were cancelled at Dusseldorf University Hospital in Germany due to a ransomware at-

tack that occurred in September 2020 [23]. In September 2022, an additional ransomware incident unfolded in Corbeil-Essonnes, which had an impact on the blood work laboratory and imaging services [30]. The University Medical Centre in Maastricht (MUMC+) was the target of a Distributed Denial of Service (DDoS) attack in January 2023 [27]. Sites affiliated with nine hospitals in Denmark were rendered inaccessible in February of that year due to DDoS attacks [27]. Early in March, Hospital Clínic de Barcelona in Spain declared a ransomware attack that rendered clinical records inaccessible, necessitating the cancellation of thousands of appointments, including non-urgent procedures and patient exams [27]. In March 2023, a ransomware attack targeted the Centre Hospitalier Universitaire Saint-Pierre in Brussels, necessitating the reliance of staff on paper records. As a precautionary measure, ambulances and medical vehicles were redirected to adjacent establishments [27]. A cyberattack was also disclosed in March 2023 by Walsall Healthcare NHS Trust, a public sector healthcare provider in the United Kingdom that serves approximately 260,000 individuals [27]. The list goes on, clearly showing the magnitude of this issue.

Indeed, these healthcare infrastructure and hospitals are typical examples of the NCP continuum, comprising highly computerized, highly interconnected services and devices, both for the purely medical technical aspects and for the management of daily operations, staff and patients. Yet, these have historically been very poorly secured against cyberattacks, making them tempting targets for cybercriminals.

In 2020, 60% of the total ransomware attacks reported in the USA were specifically directed at the healthcare sector [24]. Cybercriminals frequently targeted medical suppliers, utilizing third-party organizations to gain unauthorized access to healthcare systems and bypass internal protections. In 2021, 60% of healthcare data breaches were attributed to third-party vendors [24]. The year 2021 witnessed several significant healthcare-related data breaches, with over 40 million patient records compromised in the USA [24]. Shockingly, the prevalence of ransomware attacks on healthcare organizations soared to 66%, a substantial increase from the previous year's 34% [27]. The situation worsened in 2022, as over 50 million patient records were compromised, and a staggering total of 905 incidents were reported [24].

The situation is as bad in Europe. A comprehensive overview of the cybersecurity threats between 2019 and 2023 in the EU health sector, as well as more examples of attacks, can be found in ENISA's detailed report [31].

However, in spite of all these attacks, an August 2021 survey revealed that cybersecurity ranked as a high-priority investment for less than 11% of hospital IT executives [22]. Paradoxically, during the same period, 48% of these executives reported instances of either forced or proactive shutdowns of their systems in the preceding six months, attributable to ransomware attacks or queries. A recent report by ENISA [31] highlighted that surveys found that merely 27% of organizations surveyed within the health sector had established a dedicated ransomware defense program. Furthermore, a concerning 40% of these organizations lacked a security awareness program for non-IT staff. The shortfall in cybersecurity preparedness is further underscored by the fact that 95% of health organizations faced challenges during risk assessments, with a striking 46% having never conducted a risk analysis.

Yet, in addition to the disruption of services and risk of harm to humans, the financial costs are staggering. The average cost of a healthcare data breach reached \$9.42 million in 2021 [25]. In 2023, while the average cost of a data breach across all industries was \$4.45 million, the healthcare sector incurred the highest average cost among them, at \$10.93 million [26]. This signifies a substantial 53.3% increase in healthcare data breach costs over the past three years.

The magnitude of the issue calls for immediate and strong action. Solutions exist, and are effective. Notably, the healthcare industry demonstrated a noteworthy average cost savings of \$2 million when equipped with incident response (IR) and testing teams, in stark contrast to those without such provisions [26]. Moreover, health organizations leveraging artificial intelligence (AI) and automation in their cyber responses achieved substantial average cost savings of \$850,000. The EU is also funding research projects to protect hospitals from cyber attacks, and co-funding cybersecurity procurement for hospitals, meeting 50% of the cost of new measures [28]. These efforts must be expanded until a proper level of cybersecurity is reached in healthcare infrastructure and hospitals.

Conclusion

For the NCP to succeed, it has to elicit trust from the users. The latter hinges on a proper level of cybersecurity.

This article highlighted technical areas in which the NCP could be prone to security issues and vulnerabilities, e.g. its extended code base and communications between services, as well as NCP application domains that concretely reach the physical world, namely critical supply chains and healthcare systems and infrastructures, where cyberattacks are common and known to have dramatic effects. Cybersecurity efforts are crucial in these areas and must be intensified.

Other technical areas exist nonetheless that could not be addressed in this article, but present weaknesses and contribute to the attack surface of the NCP. For example, the IoT (Internet of Things) segment of the NCP often relies on objects that radio communicate (using wifi, Bluetooth, ZigBee and many other kinds of protocols), yet most of these devices' communications are very poorly secured and easy to hack. Another area is the human interface the NCP offers to humans, such as the browser (see the relevant article in this HiPEAC Vision 2024), or LLMs (Large Language Models) prompts. The latter offer scores of new kinds of security issues that will have to be taken into account for the LLMs segment of the NCP to succeed as well.

The NCP thus presents formidable opportunities, but these opportunities could be wasted because of cybersecurity issues if they are not properly taken care of.

References

- [1] Java Logging Package RCE Vulnerability. <https://media.cert.europa.eu/static/SecurityAdvisories/2021/CERT-EU-SA2021-067.pdf>
- [2] Log4j vulnerability - update from the CSIRT's Network. <https://www.enisa.europa.eu/news/enisa-news/log4j-vulnerability-update-from-the-csirts-network>
- [3] 2020 United States federal government data breach. https://en.wikipedia.org/wiki/2020_United_States_federal_government_data_breach
- [4] The SolarWinds Cyber-Attack: What You Need to Know. Center for Internet Security. March 2021. <https://www.cisecurity.org/solarwinds>
- [5] Actors behind PyPI supply chain attack have been active since late 2021. Dan Goodin. Ars Technica. September 1, 2022. <https://arstechnica.com/information-technology/2022/09/actors-behind-pypi-supply-chain-attack-have-been-active-since-late-2021/>
- [6] Saltzer, Jerome H.; Schroeder, Michael D. (1975). "The protection of information in computer systems". *Proceedings of the IEEE. Institute of Electrical and Electronics Engineers (IEEE)*. 63 (9): 1278–1308.
- [7] Stephen Checkoway, et al. Iago attacks: why the system call API is a bad untrusted RPC interface. *Architectural Support for Programming Languages and Operating Systems (ASPLoS) 2013*: 253-264
- [8] Jo Van Bulck, et al. A Tale of Two Worlds: Assessing the Vulnerability of Enclave Shielding Runtimes. *CCS 2019*: 1741-1758
- [9] Hugo Lefeuvre, et al. Assessing the Impact of Interface Vulnerabilities in Compartmentalized Software. *Network and Distributed System Security Symposium (NDSS) 2023*.
- [10] Aurora Generator Test. Wikipedia. (accessed 2023-12-04). https://en.wikipedia.org/wiki/Aurora_Generator_Test
- [11] Sabotage: Code added to popular NPM package wiped files in Russia and Belarus. Dan Goodin. Ars Technica. March 18, 2022. (accessed 2023-12-04). <https://arstechnica.com/information-technology/2022/03/sabotage-code-added-to-popular-npm-package-wiped-files-in-russia-and-belarus/>
- [12] Inside the Cunning, Unprecedented Hack of Ukraine's Power Grid. *Wired*. 3 March 2016. (accessed 2023-12-04). <https://www.wired.com/2016/03/inside-cunning-unprecedented-hack-ukraines-power-grid/>
- [13] Petya ransomware attacks. (accessed 2023-12-04). [https://en.wikipedia.org/wiki/Petya_\(malware\)](https://en.wikipedia.org/wiki/Petya_(malware))
- [14] The Untold Story of NotPetya, the Most Devastating Cyberattack in History. Andy Greenberg. September 2018, *Wired*. (accessed 2023-12-04). <https://www.wired.com/story/notpetya-cyberattack-ukraine-russia-code-crashed-the-world/>
- [15] Infer Static Analyzer (accessed 2023-12-04) <https://fbinfer.com/>
- [16] CodeQL (accessed 2023-12-04) <https://codeql.github.com/>
- [17] American Fuzzy Lop plus plus (AFL++) (accessed 2023-12-04) <https://github.com/AFLplusplus/AFLplusplus>
- [18] libFuzzer – a library for coverage-guided fuzz testing (accessed 2023-12-04) <https://lvm.org/docs/LibFuzzer.html>

- [19] Open Source Security Foundation (OpenSSF) Best Practices Working Group. Compiler Options Hardening Guide for C and C++. 2023-11-29 (accessed 2023-12-04) <https://best.openssf.org/Compiler-Hardening-Guides/Compiler-Options-Hardening-Guide-for-C-and-C++.html>
- [20] CompCert (accessed 2023-12-04) <https://compcert.org/>
- [21] Lacombe, G., Feliot, D., Boespflug, E. et al. Combining static analysis and dynamic symbolic execution in a toolchain to detect fault injection vulnerabilities. J Cryptogr Eng (2023). <https://doi.org/10.1007/s13389-023-00310-8>
- [22] Cyberattacks top list of 2022 health tech hazards alongside supply chain problems, damaged infusion pumps. <https://www.fiercehealthcare.com/tech/cyberattacks-top-list-2022-health-tech-hazards-ecri-report-alongside-supply-chain-problems>
- [23] Cyber attacks in healthcare: the position across Europe. Dorian Rees. 2021-06-18 (accessed 2023-12-04). <https://www.pinsentmasons.com/out-law/analysis/cyber-attacks-healthcare-europe>
- [24] Healthcare Cyber Attack Statistics 2022: 25 Alarming Data Breaches You Should Know. Joel Witts. 2023-03-28 (accessed 2023-12-04). <https://expertinsights.com/insights/healthcare-cyber-attack-statistics/>
- [25] The Average Cost of a Healthcare Data Breach is Now \$9.42 Million. Steve Alder. 2021-07-19. (accessed 2023-12-04) <https://www.hipaajournal.com/average-cost-of-a-healthcare-data-breach-9-42-million-2021/>
- [26] Cost of a data breach 2023: Healthcare industry impacts. Michelle Greenlee. 2023-08-16. (accessed 2023-12-04) <https://admin05.dev.blogs.cis.ibm.net/articles/cost-of-a-data-breach-2023-healthcare-industry-impacts/>
- [27] Ransomware Is a Growing Threat for European Healthcare Organisations. Jason O'Connor. 2023-04-21. (accessed 2023-12-04) <https://www.keepersecurity.com/blog/2023/04/21/cyberattacks-soar-across-the-european-healthcare-sector/>
- [28] The race to make hospitals cybersecure. Tom Cassauwers. 2023-05-24. (accessed 2023-12-04) <https://ec.europa.eu/research-and-innovation/en/horizon-magazine/race-make-hospitals-cybersecure>
- [29] Cyber attacks hit two French hospitals in one week. 2021-02-16. (accessed 2023-12-04) <https://www.france24.com/en/europe/20210216-cyber-attacks-hit-two-french-hospitals-in-one-week>
- [30] Paralysed French hospital fights cyber attack as hackers lower ransom. 2022-09-22. (accessed 2023-12-04) <https://www.rfi.fr/en/france/20220902-paralysed-french-hospital-fights-cyber-attack-as-hackers-lower-ransom-demand>
- [31] Health Threat Landscape. ENISA. 2023-07-05. (accessed 2023-12-04) <https://www.enisa.europa.eu/publications/health-threat-landscape?v2=1>

Bart Coppens is a part-time assistant professor and a post-doctoral researcher in the Electronics department of Ghent University, Ghent, Belgium.

Olivier Zendra is a Tenured Computer Science Researcher at Inria, Rennes, France.



In the Next Computing Paradigm (NCP), numerous home sensors will communicate with servers and services. The multitude of data and communication is bound to raise privacy issues and these must be taken care of for the NCP to succeed.

More data for the NCP implies more privacy risks

by Bart Coppens and Olivier Zendra

Key insights

- Even the most innocuous data sources, such as those from IoT-based sensors, can be used to infer personal and private information.
- Integrating different sources of data significantly increases the amount of personal and private information that can be inferred.
- When such data is exposed, this creates an enormous privacy problem for the affected people.
- Systems that are tightly integrated into people's daily lives thus lead to a significant privacy risk.
- The NCP is such a tightly-integrated system which could lead to a significant privacy risk if not properly designed and implemented.

Key recommendations

- Promote research into technologies that enhance people's privacy and reduce the risks and impact of leaks of private data.
- Stand by EU principles of privacy for its citizens, requiring companies to actively adhere to the principles of privacy by design. In particular, functional requirements for systems should include:
 - be designed to not leak data in any form whatsoever, except with the explicit consent of the users; no backdoors should be allowed.
 - give the user the option to decide (with explicit consent) what data is collected, and to understand what that consent implies.

Innocuous Data Sources Can Leak Personal and Private Data

Data stored by systems can come from numerous sources. One source is users uploading or explicitly sharing their own data (when writing in an online editor, sharing photos, sending messages, etc.). In these cases, users often know that they share information with others and can mentally distinguish between personal and private data. However, this explicitly user-shared data is only the tip of the proverbial iceberg of data. Other sources of data include the (logs of the) user's activity, and the sensors that increasingly pervade our daily lives. Users are typically unaware of these sources and cannot estimate how personal or private data extracted from them can be. In this article, we focus on the privacy of data inferred from actual, physical sensors; however data extracted from digital interactions with a system presents major privacy concerns too.

Different sensors keep being introduced into our lives. Obvious ones include security cameras in our cities and villages, and ANPR cameras that track car license plates. Our mental picture of these cameras seems quite clear: they are on fixed spots, and relatively visible, and we 'know' their purpose is catching criminals. However, the privacy implications might not always be immediately as clear. But these cameras of course do not magically only turn themselves on when a convicted criminal passes by: they are always on.

While such cameras can already lead to some privacy worries, they are in public locations, and we are typically aware of them. Mobile device cameras are more intriguing. Google Glass faced a backlash, with people calling users using Glass in a creepy manner 'glassholes' [19], and venues banning their use because of privacy concerns [20,21,22]. Google Glass was first rebranded to focus on productivity in professional settings, such as technicians and factory workers [6], but was eventually completely discontinued [16].

A more nuanced example of cameras in private places are those found on laptops and smartphones. Many laptops have a front-facing camera for video calls. Unfortunately, this means the laptop camera faces the user when making video calls. This would be fine if the camera just recorded what the user wants, when the user wants it. However, an attacker could illegally access the laptop, enable the camera (which is always there) without consent or awareness of the

user, and record whatever private and personal is happening in front of the camera. Because this worries customers, some laptop makers add a privacy screen for laptop cameras so users can physically prevent unwanted recording. Since privacy screens physically obstruct the camera lens, users can clearly see when they're active. More importantly, this cannot be overridden from software. The latter is crucial: if an attacker has gained control of the laptop and secretly enabled the camera, the attacker could also try to override any software-based protection. Note that not all users care about this issue. Some users will trust that their browser's / smartphone's webcam prompts are sufficient protection, others will say they don't think they'll be targeted, or aren't high-profile enough to warrant circumventing protections, and others will say they don't care if they're filmed by an attacker, or that the privacy screen's hassle outweighs its benefits. An additional consideration is that if an attacker takes over a device to control a camera, that attacker probably has access to all files on that device. While people can know upfront what information is (not) stored on a device, they cannot know what will be said or done in front of a camera. It is important to realize that the aforementioned considerations divide people: some find them unimportant, while others find them crucial. These very personal and individual trade-offs have varied outcomes for various persons. This is fine because such users were given the choice, can weigh the advantages and downsides *for their specific situation*, and can make an informed decision.

However, given the success of these privacy screens, this example should be considered more broadly. While users are aware and actively decide whether or not their laptop's camera can record them, they are typically less aware of their *smartphone* cameras. These however may partially face the same scenes that the laptop's webcam was deemed too risky or private to capture. People rarely even *think* about their phones' cameras, let alone privacy screens for them. Even when people mask their laptop's camera, their laptop and smartphone contain microphones able to record private talks. As with cameras, some individuals trade off that this is not a real threat, while others do.

Microphones may indeed constitute a privacy risk. Academics were (and still are) reluctant to provide recordings of online classes due to privacy concerns [24] and the risk of excerpts being

taken out of context [23], even though such classes are semi-public. In intimate conversations, privacy becomes more obvious: who wants an automated assistant listening in on their MD or partner in bed? While some might not care, others would. This is similar to Orwell's "1984" novel, where telescreens gather video and audio [25], except that now it is primarily corporations doing so (although some governments still strive to get that information).

Once pointed out, most people can reason about the privacy implications of cameras and microphones spying on private scenes and conversations. But not all sensors and types of data will elicit the same reaction, because even though the privacy implications are high, it can be harder to intuitively grasp attackers' imaginative possibilities. For example, not only may microphones be used to listen to conversations, but an eavesdropper can recognise the sounds a keyboard makes when typing, which can reveal confidential messages or passwords [26,27]. A smartphone could thus even leak information typed on a non-internet-connected computer. A microphone isn't even required for such an attack: an accelerometer in a smartphone on the same table as a laptop can capture what's being typed on the laptop through table vibrations [17].

One last example: smart building CO₂ sensors. These can monitor building air quality and control ventilation. However, these sensors can detect a room's occupant [18].

Again, these sensors being present does not mean data *will* be utilised against user privacy expectations. But there is a non-zero *risk* that an attacker will surreptitiously enable sensors when they should not be, or that sensors' data will be exploited to deduce and disclose personal information. How much risk is tolerated varies by user.

Integrating Different Data Sources Increases Privacy Issues

More sensors and data sources increase privacy danger. The main reason is that all of these sources have a privacy risk, and having more combines them. However, that is not the end of the story.

Multiple data sources that *individually* do not reveal privacy-sensitive information may *leak* it

when *combined*. Consider GPS position. As an individual data source, it already allows attackers to determine that people's homes are where their GPS signals remain most nights, their workplaces are where they stay during the day, etc. GPS combined with *public map data* reveals vast amounts of private information. A GPS location staying near churches, gay clubs, abortion clinics, etc., can reveal very sensitive information about religion, sexual preferences, health, etc. People may be wary of sharing such precise location data with huge firms like e.g. Google, but they often forget that when using Google Maps or Waze they transmit the same information to Google.

One example is de-anonymizing (potentially innocent) media-reported crime suspects. This reporting can include initials, age, profession, location, etc. While each of these bits of information is *individually* shared harmlessly with many people, *combined* they deanonymize reports, identifying the report's person. This information can be further integrated with semi-public sources like white pages, data breaches, and leaks [3].

Fitness monitoring social networks like Strava are another example. Some people want to share their cycling and running successes, but they don't want others to know where they live or work. Sharing their entire track would however reveal that. This caused social media platforms to implement endpoint privacy zones that hide track ends. However, these websites reported the track length and middle part of the track. This information can be used with maps to confine the start and end positions of the tracks, revealing critical information [4]. Similarly, soldiers' exercise paths have been used to locate military bases abroad, raising concerns about location data shared with fitness trackers like Strava [2]. Using runners' itineraries, Figure 1 even shows a military base's internal map, in Helmand Province, Afghanistan [2].



Figure 1. A military base in Helmand Province, Afghanistan with routes taken by joggers highlighted by Strava. Photograph: Strava Heatmap. Caption & image from The Guardian [2].

This also ties in with the difficulty of anonymization of data sets. If one collects precise GPS information, but ‘anonymizes’ it by removing the name, it is really not anonymized in any practical sense. The more information can be combined, the more options to deanonymize, the more options to reveal privacy-sensitive information [28].

The Risks of Systems that Tightly Integrate with People’s Lives

What does this mean for systems that tightly integrate into people’s lives? We discussed smartphones, computers, fitness trackers, etc., and their sensors before. The privacy risks of all the data such technologies can collect are obvious. Perhaps surprisingly, cars too contain multiple sensors that are tightly integrated into our daily lives. Systems are being developed to measure driver eye activity, tension, and well-being [9]. The Mozilla Foundation examined car manufacturers’ privacy practices [1,7]. The outcomes were appalling. Modern cars come with numerous sensors inside and outside and can connect to your phone. They can monitor your music, access your contact list, even record and intercept text communications [8]. This leads to privacy policies relating to music preferences, employment, and sometimes ‘sexual activity’, ‘sex life’, ‘psychological trends’, ‘intelligence,’ ‘genetic characteristics’, and many other characteristics that many people would find creepy to think their car can collect or infer [1,7,10]. These policies also explain how car companies can not only collect such data, but also sell them [1,7,10].

It should be clear that systems which tightly integrate in our daily lives can collect a dizzying and overwhelming amount of private data. Once collected, what can be done with the data? Of course they can be sold. Even when they are not sold, but are merely collected for ‘internal’ or ‘business’ purposes, these sensitive data persist as a privacy risk. Who has access to such data is unclear even in such cases. A few years ago, for Siri’s voice assistant quality control, Apple supplied contractors portions of recorded conversations, even those where Siri misidentified its activation instruction [13]. These fragments included recordings of people having sex, or discussing confidential medical information [13]. Another, worse example is Tesla personnel internally exchanging and joking about embarrassing or intimate customer recordings collected by Tesla cars [11].

Even when all employees and contractors behave, a breach could expose this data, since automotive manufacturers are also involved in data leaks and breaches [12,14,15].

Even if not all such information would be explicitly kept or inferred for European customers due to regulations, it’s important to realize that *the sensors are still there*. The sensors will still be there, recording data. Even if not everything is *sent* or *stored* by default, it’s still possible to *enable* recording or storing of data. The data, once recorded, still allow for inferences about privacy-sensitive topics to be made, if not now, then perhaps later if it is stored. The sensors, even if they *should* be disabled and not recording, are not necessarily so. For example, consider the Pegasus spyware, which amongst its many spying features allows a remote attacker to surreptitiously activate cameras and microphones on an infected smartphone [5]. This spyware was used by authoritative governments, allowing them to spy not only on their own citizens, but also on foreign citizens [5]. Importantly, if governments are able to do so, criminals can try to make use of the same mechanisms to spy on people. This should be a dire warning to *not* require (and even *forbid*) such backdoors, as they also introduce a significant additional risk to the privacy of end users because that back door could also enable criminals to listen in on them directly.

Some of the above privacy risks may surprise many people. Often individuals are ignorant of the sensors and the data they record or infer, alone or in combination. They are often unaware that talking about sensitive information near a seemingly-inactive phone might still be sending fragments of that conversation to contractors or that their car might take nude pictures of them and send them to the car manufacturer who can then make fun of them. They are often unaware that their car’s GPS (built-in or smartphone app) can deduce their sexual preference, pregnancy, health, etc.

Complex systems with many sensors always require a balance: the more you want a system integrated into people’s life to measure, the more sensors it needs, and the greater the privacy risk. Not everyone has this issue, everyone’s risk assessment trade-off is different. Users should be able to determine their own trade-offs and *accept risk* if relevant. But to do so, ***the user must be aware of the risk***. Too often, privacy policies hide these risks in tiny corners with opaque legal wording

and/or imprecise language, obscuring the actual impact. Users can authorise or decline data collection only if they understand it. Furthermore, their **consent must be a meaningful choice**, not simply 'either use our product with all sensor data being collected, or don't use it at all'. Expecting customers to either have their car collect mental state data or not drive a (modern) car is a false choice. Users must have a **meaningful choice to disable unnecessary sensors and collected and inferred data without losing the entire system access**. In addition, systems should **never disclose or leak data without the explicit, informed and meaningful consent of the user**.

Finally, it is worth considering the design space of possible approaches to fulfil such goals. Webcam privacy screens are an interesting design choice, though not always applicable. It plainly reveals whether explicit or accidental video recording is allowed, in a way that is not corruptible or subvertible in software. This visual cue makes the webcam difficult to overlook. The laptop's non-webcam functions can still be used when the webcam is disabled. Thus, a system in which the sensor whose functionality a user does not need can be *physically* removed (either temporarily or permanently, e.g. by physically disabling or removing it) can give the user a clear *choice* that is not subvertible by software, hence *reducing the attack surface* in a meaningful way.

To achieve such goals, research into technologies that improve privacy and prevent data leaks is crucial. This should cover both the *design* of systems, as well as their *implementation*.

The NCP is precisely such a system integrating tightly into people's lives, comprising many sensors, and combining many different data sources. From a privacy perspective, it is crucial that its design includes the above **requirements** for informing the user clearly about the potential risks, allowing the user to enable or disable certain sensors and information sources from being integrated into the system, and only sharing information with *any* party with the user's **explicit, informed and meaningful consent**, so that each user can make their own trade-off. Again, the NCP's design and implementation must both meet privacy criteria. Its functionalities must gracefully degrade depending on user privacy choices.

Conclusion

Even innocuous data sources like IoT sensors can reveal personal information. Integrating data sources greatly increases the amount of personal and private information inferred. When such data is exposed, affected users face huge privacy issues. Integrating systems tightly into people's daily life increases privacy risks. Due to its intimate integration, the NCP could present a considerable privacy risk if not properly designed and implemented. Thus, the NCP must be designed and implemented to let users consent — or deny — to data collection and data sharing knowingly. In addition, the NCP must make that choice a meaningful one, by having only the specific features, in the narrowest sense, that actually need the denied information degraded by that refusal.

References

- [1] It's Official: Cars Are the Worst Product Category We Have Ever Reviewed for Privacy. Jen Caltrider, Misha Rykov and Zoë MacDonald. Privacy Not Included, Mozilla. 2023-09-06 (accessed 2023-11-23) <https://foundation.mozilla.org/en/privacynotincluded/articles/its-official-cars-are-the-worst-product-category-we-have-ever-reviewed-for-privacy/>
- [2] Fitness tracking app Strava gives away location of secret US army bases. Alex Hern. The Guardian. 2018-01-28 (accessed 2023-11-27). <https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases>
- [3] De Boeck, Kevin, et al. "Poster: The impact of public data during de-anonymization: a case study." 7th IEEE European Symposium on Security and Privacy, Date: 2022/06/06-2022/06/10, Location: Genoa. 2022.
- [4] Dhondt, Karel, et al. "A Run a Day Won't Keep the Hacker Away: Inference Attacks on Endpoint Privacy Zones in Fitness Tracking Social Networks." Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. 2022.
- [5] Stephanie Kirchgaessner, Paul Lewis, David Pegg, Sam Cutler, Nina Lakhani and Michael Safi. "Revealed: leak uncovers global abuse of cyber-surveillance weapon". The Guardian, 2021-07-21 (accessed 2023-11-23). <https://www.theguardian.com/world/2021/jul/18/revealed-leak-uncovers-global-abuse-of-cyber-surveillance-weapon-nso-group-pegasus>
- [6] Google Glass 2.0 Is a Startling Second Act. Steven Levy. Wired. 2017-07-18 (accessed 2023-11-23) <https://www.wired.com/story/google-glass-2-is-here/>
- [7] What Data Does My Car Collect About Me and Where Does It Go? Jen Caltrider, Misha Rykov and Zoë MacDonald. Privacy Not Included, Mozilla. 2023-09-06 (accessed 2023-11-23) <https://foundation.mozilla.org/en/privacynotincluded/articles/what-data-does-my-car-collect-about-me-and-where-does-it-go/>
- [8] Court rules automakers can record and intercept owner text messages. Suzanne Smalley. The Record. 2023-11-08 (accessed

- 2023-11-23) <https://therecord.media/class-action-lawsuit-cars-text-messages-privacy>
- [9] Harman's driver-monitoring system can measure your heart rate. Rebecca Bellan. TechCrunch. 2023-01-04, accessed 2023-11-23 <https://techcrunch.com/2023/01/04/harmans-driver-monitoring-system-can-measure-your-heart-rate/>
- [10] Nissan. Privacy Not Included, Mozilla. 2023-08-18 (accessed 2023-11-23) <https://foundation.mozilla.org/en/privacynotincluded/nissan/>
- [11] Tesla workers shared sensitive images recorded by customer cars. Steve Stecklow, Waylon Cunningham and Hyunjoo Jin. Reuters. 2023-04-06 (accessed 2023-11-23) <https://www.reuters.com/technology/tesla-workers-shared-sensitive-images-recorded-by-customer-cars-2023-04-06/>
- [12] More than 2 million Toyota users face risk of vehicle data leak in Japan. Daniel Leussink and Kantaro Komiya. Reuters. 2023-05-12 (accessed 2023-11-23) <https://www.reuters.com/business/autos-transportation/toyota-flags-possible-leak-more-than-2-mln-users-vehicle-data-japan-2023-05-12/>
- [13] Apple contractors 'regularly hear confidential details' on Siri recordings. Alex Hern. The Guardian. 2019-07-26 (accessed 2023-11-23) <https://www.theguardian.com/technology/2019/jul/26/apple-contractors-regularly-hear-confidential-details-on-siri-recordings>
- [14] Volkswagen, Audi disclose data breach impacting over 3.3 million customers, interested buyers. Charlie Osborne. ZDNet. 2021-06-14 (accessed 2023-11-23) <https://www.zdnet.com/article/volkswagen-audi-disclose-data-breach-impacting-over-3-3-million-customers-interested-buyers/>
- [15] 1.6 million hit in possible Mercedes-Benz data breach — what you need to know. Paul Wagenseil. Tom's guide. 2022-10-20, (accessed 2023-11-23) <https://www.tomsguide.com/news/mercedes-benz-data-breach>
- [16] RIP (again): Google Glass will no longer be sold. Samuel Axon. Ars Technica. 2023-03-16 (accessed 2023-11-23) <https://arstechnica.com/gadgets/2023/03/google-glass-is-about-to-be-discontinued-again/>
- [17] Marquardt, Philip, et al. "(sp)iPhone: Decoding vibrations from nearby keyboards using mobile phone accelerometers." Proceedings of the 18th ACM conference on Computer and communications security. 2011.
- [18] da Silva, Marlon P., et al. "Impact of using a privacy model on smart buildings data for CO2 prediction." arXiv preprint arXiv:2306.00766 (2023).
- [19] Google Glass advice: how to avoid being a glasshole. Samuel Gibbs. The Guardian. 2014-02-19 (accessed 2023-11-27) <https://www.theguardian.com/technology/2014/feb/19/google-glass-advice-smartglasses-glasshole>
- [20] Theater chain bans Google Glass. David Kravets. Ars Technica. 2014-06-10 (accessed 2023-11-27) <https://arstechnica.com/tech-policy/2014/06/theater-chain-bans-google-glass/>
- [21] Unbearable wearable: Bar bans Google Glass, boots 'rude' user. Devin Coldewey. NBC News. 2013-11-26 (accessed 2023-11-27) <https://www.nbcnews.com/technology/unbearable-wearable-bar-bans-google-glass-boots-rude-user-2d11660837>
- [22] San Francisco bar bans Google Glass for fear of secret recordings. Gale Holland. Los Angeles Times. 2014-03-04 (accessed 2023-11-27) <https://www.latimes.com/local/lanow/la-me-ln-bar-bans-google-glass-wearers-20140304-story.html>
- [23] Is This Thing On? Jeffrey Aaron Snyder and Amna Khalid. Inside Higher Ed. 2020-10-13 (accessed 2023-11-27) <https://www.insidehighered.com/views/2020/10/14/thorny-issues-surrounding-classroom-recordings-reasonable-accommodations-and>
- [24] Should Professors Still Record Lectures? Maybe. Maybe Not. Susan D'Agostino. Inside Higher Ed. 2022-09-06 (accessed 2023-11-27) <https://www.insidehighered.com/news/2022/09/07/should-professors-still-record-lectures-maybe-maybe-not>
- [25] 1984. George Orwell. Published 1949-06-08 by Secker & Warburg.
- [26] Asonov, Dmitri, and Rakesh Agrawal. "Keyboard acoustic emanations." IEEE Symposium on Security and Privacy, 2004. Proceedings. 2004. IEEE, 2004.
- [27] Zhuang, Li, Feng Zhou, and J. Doug Tygar. "Keyboard acoustic emanations revisited." ACM Transactions on Information and System Security (TISSEC) 13.1 (2009): 1-26.
- [28] Data anonymization in Big Data scenarios: an open challenge to become GDPR compliant. Sara El Kortbi Martínez. Gradiant. 2021-11-11 (accessed 2023-11-29) <https://www.gradiant.org/en/blog/infinitech-data-anonymization-big-data-gdpr/>

Bart Coppens is a part-time assistant professor and a post-doctoral researcher in the Electronics department of Ghent University, Ghent, Belgium.

Olivier Zendra is a Tenured Computer Science Researcher at Inria, Rennes, France.



Currently, and in the Next Computing Paradigm (NCP) as well, one of the major ways of interacting with the Web is through the browser. But tracking users through their browser is commonplace, and often underestimated, with the user not being aware of them. The browser is thus a keystone to user privacy in the NCP.

The browser: the key to your privacy on the Web

by Walter Rudametkin and Olivier Zendra

Key insights

- User tracking through the browser is extremely widespread. Users are tracked almost all the time, on most websites.
 - first-party and third-party cookies are used almost all the time
 - browser fingerprinting is used less often but is powerful and complementary to cookies
- IP address tracking, combined with fingerprinting, is a very effective for replacing third-party cookies
- Tools exist to check that user preferences for 3rd-party cookies are respected, but they are not easily available to end-users.
- User preferences for controlling fingerprinting are non-existent, whereas browser fingerprinting falls into the same GDPR data category as cookies and other trackers
- Browser fingerprinting has security uses that websites actively use
- Vendors are adding new functionalities to browsers at a frenetic pace, not enough effort is being put into minimizing the privacy risks

Key recommendations

- Promote research into technologies and tools that break the uniqueness of people when browsing, hence preserve their privacy.
- Promote 3rd-party cookies interdiction by the browsers.
- Promote research into technologies and tools to identify, display and allow users to limit fingerprinting techniques.
- Enforce GDPR wrt. user tracking, ensuring that the collected data is strictly necessary to the usages accepted by the users. This should include all tracking technologies: cookies, IP tracking, fingerprinting...

Browser tracking, for good or for bad, it's there

Modern Internet technologies are progressing at an amazing rate and redefining the limits of the Internet. Some research in the field of cybersecurity focuses on identifying new security and privacy threats on the Web, with extensive expertise in an identification technique that can be exploited for security and tracking purposes, called browser fingerprinting. In this article, we consider two facets. The first attempts to improve authentication and explores the use of browser fingerprinting to improve the security of websites and user accounts by verifying elements of their device's configuration. The second explores how novel features being quickly added to browser's can be exploited to enhance fingerprinting and track users without their knowledge, and also tries to understand how different tracking technologies combine into persistent, omnipresent tracking mechanisms. This is particularly important as cookie-based tracking, which is very prevalent and based on third-party cookies, is inevitably going to be being deprecated [1], sometime in the future (Chrome has pushed the deadline to late 2024 [17]), while browser fingerprinting is potentially "an invasive workaround to replace cookies" [2], and Google, who controls development on the Chrome browser and derivatives [3], exerts pressure to control the future of tracking [4].

The ability to track users on the Internet and their online habits is indeed lucrative to content producers and advertising companies, such as major Big Tech players, as well as being very intrusive to user privacy. Furthermore, users have grown accustomed to "free" services that provide ever more extensive and impressive functions, everything from video services, social media platforms, to video games, office platforms, and much more is now all Web-based. In essence, the Web browser is our window to the Internet and the NCP, our window to the world. Yet, these services are anything but free; aggressive tracking and profiling is pushed for monetization. Studies have shown that user tracking continues to increase on popular websites [7], [8]. State-of-the-art tracking techniques assign a unique identifier, which is stored in the browser—either as a cookie or some other storage mechanism (e.g., local storage, Etags). To protect users, private browsing modes and extensions automatically delete cookies and clear storages at the end of a session, decreasing the efficiency of the standard cookie-based tracking techniques.

But to compensate for this process of deleting cookies and blocking stateful tracking, a new identification technique that leaves no traces on the user's device is being used. It is called Browser fingerprinting [11], [12]. *Browser fingerprinting* is the process of identifying devices by accessing a collection of relatively stable attributes through Web browsers. We call the identifiers browser *fingerprints*. Fingerprints are stateless; no information is stored on the client's device. Browser fingerprinting exploits the diversity of modern web configurations, technologies, protocols and APIs (Application Programming Interfaces) to uniquely identify devices. And contrary to tracking cookies that are stored on the device and can be erased, fingerprints are stored on servers the user has no control over. Encryption does little to limit fingerprinting because it is performed by the website you visit; it is not a sniffing nor man-in-the-middle attack.

Extensive studies of browser fingerprinting, including those from the *Am I Unique* platform [27] for the past 10 years, have identified three main properties of browser fingerprinting that make it both a risk to privacy, but also useful for security. The first property, **uniqueness**, is the power to uniquely identify a device. Fingerprint uniqueness, although not a perfect identifier, has statistical qualities that allow uniquely identifying a high percentage of both desktop computers and mobile devices [12]. The second property, **linkability**, is the capacity to re-identify, or *link*, fingerprints from the same device over time. This is arguably the main risk to privacy and enables *fingerprint tracking*. Some devices are highly trackable, while other devices' fingerprints are too similar to be tracked [9]. The third and final property is **consistency**, which refers to the capacity to verify the attributes in a fingerprint. Through redundancies, correlations or dependencies, many attributes are verifiable, making them difficult to spoof convincingly. Most countermeasures to browser fingerprinting are identifiable through inconsistencies [10], a useful property for security applications, but also increases the privacy risks for users.

These are real concerns and the use of browser fingerprinting for security could be a net positive to companies, to universities and to society. However, abuses may arise. Finding a good balance between security, usability and the risks to privacy is a major and fundamental challenge to get a better, safer, more privacy friendly internet. This is important to avoid social cooling, whose concept is simple: *if you feel you're being watched, you change your behavior*, which is, in essence, a form of restricting

your freedom. Among the things that change, are what you say and how you act. Through what is known as surveillance capitalism, many people are discovering that everything we do is being monitored, dissected and monetized.

This article addresses two main aspects that are both complementary yet opposing by nature: browser identification for security purposes, and browser identification for tracking purposes. The first aspect will focus on **i) browser fingerprinting for Web authentication**. The main objective is to enhance and augment multi-factor authentication through advanced browser fingerprinting. This requires identifying APIs that can be used to create hard-to-forge browser fingerprints and authentication algorithms that resist different attack models. The second aspect will focus on **ii) stateless tracking techniques** that create privacy risks in the browser. Since browser fingerprinting is rarely used by itself, we'll explain how it complements other tracking techniques like IP addresses and cookie-based tracking.

Browser fingerprinting for authentication

Advanced browser fingerprinting can be used as a configurable authentication mechanism. It even has the potential to be the only authentication mechanism when used in very low-security, public websites on the Internet. It can be used to block bots [6] and other fraudulent users. It also has the potential to be an additional security factor in Multi-Factor Authentication (MFA) schemes. Besides strengthening a session's initial authentication, it can be used for continuous session authentication [13] to protect against session hijacking attacks [14]. In many contexts fingerprinting is fully transparent to users, and unlike security cards, code generating keys, apps, SMS verification codes, users do not have to do anything to improve their security. In more restricted contexts, administrators can even enforce different policies, such as enrolling fingerprints from devices that connect from trusted IP addresses (e.g., in an internal enterprise network), and then verifying these fingerprints when the same users connect from untrusted IP addresses. Plugging the browser fingerprinting authentication process into existing authentication systems raises issues, such as fingerprint forgeability, usability and effectiveness. These could be addressed with attributes that

focus on identifying hard-to-forge hardware characteristics of the device and by designing dynamic challenge-response tests to limit replay attacks.

But building browser fingerprints that are beneficial to authentication raises issues. While modern browsers are strengthening the protection of their users' privacy by deploying defenses against potential privacy leaks, the evolution of Web programming technologies pushes towards personalization that inevitably leads to the disclosure of sensitive parameters. This continuous tension between technological evolution and browser engagement stresses browser fingerprinting techniques to exploit a moving target set of fingerprintable attributes. Moreover, websites that wish to use fingerprinting to improve the security of their users and services, e.g. as a second factor authentication, must decide what attributes to collect, yet, since fingerprints are calculated on the client's device, unforgeable attributes that resist against motivated attackers that can intercept client requests and replay their fingerprints are needed.

Furthermore, online fraudsters, malicious users and crawlers represent a growing share of Internet traffic. State-of-the-art protections rely on analyzing the traffic to detect suspicious patterns and fallback to CAPTCHAs [18] to block unwanted activity [19][20][21][22]. However, these techniques offer only partial protection at the cost of expensive monitoring. They tend to be slow, ineffective to distributed crawlers, and miss much of the actual fraudster traffic.

Websites are also strengthening their authentication processes to ensure only valid users can authenticate a session. To increase security, Multi-Factor Authentication (MFA) combines factors (e.g. credentials, SMS, smartphone apps) to prevent credentials from being used maliciously. However, these additional factors add cost (e.g., purchasing USB keys for all users) and impact the user experience (by forcing interactions with multiple devices to authenticate), which limits their adoption. Browser fingerprinting can mitigate many of these issues because a fingerprint can be connected before accessing content or an account, and can immediately detect fraudsters, in one step. Furthermore, users don't have to bear any costs and, in the case of legitimate users, this is transparent as long as the false positives are low.

Browser tracking: you're just a little different but that's enough for them to know who you are

Data collection from users of the Am I Unique browser extensions show that minor fabrication differences in GPUs can be used to uniquely identify devices despite being otherwise identical hardware, similar to research on clock differences [15]. Users who customize their computers or browsers, or their different tools, have fingerprints that are more unique than those that don't. In some cases, users are confronted with a paradox: do I modify my configuration to protect my privacy from, for example, cookies? Should I move to a privacy friendly operating system and browser? Usually this would be the obvious choice and will work for some techniques, such as blocking your cookies, but when done poorly, this can lead to differences in the browser fingerprint making you more identifiable. Furthermore, moving from popular technological choices to less popular but more privacy protecting has the same effect of increasing your browser fingerprint uniqueness. Interestingly, even in the case of ad blockers, which increase their privacy in general, could in fact make you more identifiable through fingerprintable customizations, despite the overall privacy advantages of the tool.

Most uses of browser fingerprinting for tracking purposes occur in conjunction with other techniques, such as cookie-based tracking, invisible pixels, IP address tracking, the collection of personal identifiable information (PII), and many others. This leads to a dire situation for privacy on the Web. Browser fingerprinting is a far from perfect tracking technique, but combined with cookie identifiers, IP addresses, ETags, invisible pixels, PII, and many others, it's much worse for privacy. It's indeed difficult to properly isolate tracking techniques, such as browser fingerprinting, especially when the identifiers are collected and stored server-side. IPv4 address retention is a serious problem [16], and combining this with browser fingerprinting is likely to be very effective for tracking. For example, to avoid tracking, if users erase all their cookies, they would simultaneously need to clear all caches, change their browser fingerprint, change their IP address (e.g., restart their router and obtain a new DHCP address from their ISP or use a VPN), not provide any PII, to have any strong assurance that they are not potentially and immediately reidentified. Furthermore, with third party cookies being deprecated, these techniques risk being further exploited and novel

tracking techniques, such as server-side tracking, developed and perfected. This is not a theoretical approach, Meta already provides a new server-side tracking product called the Conversions API [23] where websites can collect IP addresses, user agents, and personal identifiers and send them to Meta. Moreover, the "best practices" call for combining the Conversions API with their client-side tracker [24], the Meta Pixel [25], that now supports both first and third party cookies. You read this right, first and third party cookies, fingerprinting, IP addresses and PII are all combined to track users across any domain that uses Meta's advertising products. And Google has its own server-side product that is more extensible and configurable than Meta's called server-side tagging [26].

This begs the question, how do people effectively defend against a barrage of privacy invasive attacks? The simple answer is that there are no simple answers.

Browsers and their features are rapidly developed, with little interest or caution to privacy issues, opening the door to fingerprinting and other side-channel attacks. Browsers are very configurable and different configurations can introduce many issues. Furthermore, browsers are extensible through extensions, and these extensions can introduce privacy bugs of their own that make users identifiable (e.g., in FP-Scanner [10] we showed privacy focused extensions to often be counter-productive despite their promises). This field of research is of even more importance now that a few major actors on the Web – all of Big Tech are outside of the EU – have concentrated control over entire swaths of the Web, and have begun to push to a more closed, more tracked, more controlled and privacy destructive future. EU researchers and regulators need to step up and attack these issues from all angles, legal, technical, economical and societal.

Conclusion

The browser is one of the most important user interfaces to the NCP, that is used daily by billions of people without any second thoughts. However, browser tracking techniques can both be used for ease of use, or cause tremendous privacy issues. It is thus crucial that research and regulation address the issues of tracking, provide the users with legal rights and the technical means to verify them, including easy to use tools to detect, control, verify and limit undesirable tracking in all its forms.

References

- [1] Building a more private web: A path towards making third party cookies obsolete. Chromium Blog. 2020-01-14 (accessed 2023-11-28). <https://blog.chromium.org/2020/01/building-more-private-web-path-towards.html>
- [2] Building a more private web. Justin Schuh. 2019-08-22 (accessed 2023-11-28). <https://blog.google/products/chrome/building-a-more-private-web/>
- [3] Browser Market Share Worldwide. statcounter GlobalStats. (accessed 2023-11-28). <https://gs.statcounter.com/browser-market-share>
- [4] Google's FLoC Is a Terrible Idea. EFF. Bennett Cyphers. 202-03-03 (accessed 2023-11-28). <https://www.eff.org/deeplinks/2021/03/googles-floc-terrible-idea>
- [5] Big Tech. Wikipedia. (accessed 2023-11-28). https://en.wikipedia.org/wiki/Big_Tech
- [6] A. Vastel, W. Rudametkin, R. Rouvoy, X. Blanc. FP-Crawlers: Studying the Resilience of Browser Fingerprinting to Block Crawlers. MAD-Web'20 - NDSS Workshop on Measurements, Attacks, and Defenses for the Web. <https://hal.archives-ouvertes.fr/hal-02441653>
- [7] G. Acar, M. Juarez, N. Nikiforakis, C. Diaz, S. Gurses, F. Piessens, and B. Preneel, "FPDetective: Dusting the web for fingerprinters," Conf. on Computer and Communications Security (CCS '13). <https://www.esat.kuleuven.be/cosic/publications/article-2334.pdf>
- [8] S. Englehardt and A. Narayanan, "Online tracking: A 1-million-site measurement and analysis," CCS '16. http://randomwalker.info/publications/OpenWPM_1_million_site_tracking_measurement.pdf
- [9] A. Vastel, P. Laperdrix, W. Rudametkin, R. Rouvoy. FP-STALKER: Tracking Browser Fingerprint Evolutions. IEEE Symp. on Sec. and Privacy (IEEE S&P'18). <https://hal.inria.fr/hal-01652021>
- [10] A. Vastel, P. Laperdrix, W. Rudametkin, R. Rouvoy. FP-Scanner: The Privacy Implications of Browser Fingerprint Inconsistencies. USENIX Security 2018. <https://hal.inria.fr/hal-01820197>
- [11] P. Eckersley. How unique is your web browser? International Conference on Privacy Enhancing Technologies (PETS'10). <https://panoptickick.eff.org/static/browser-uniqueness.pdf>
- [12] P. Laperdrix, W. Rudametkin, B. Baudry. Beauty and the Beast: Diverting modern browsers to build unique browser fingerprints. IEEE Symposium on Security and Privacy (S&P'2016). <https://hal.inria.fr/hal-01285470v2>
- [13] D. Preuveneers, W. Joosen. SmartAuth: dynamic context fingerprinting for continuous user authentication. Symposium on Applied Computing SAC'15. <https://dl.acm.org/citation.cfm?id=2695908>
- [14] A. Durey, P Laperdrix, W Rudametkin, R. Rouvoy. "FP-Redemption: Studying Browser Fingerprinting Adoption for the Sake of Web Security." DIMVA'21. <https://hal.inria.fr/hal-03269174>
- [15] I. Sanchez-Rola, I. Santos, D. Balzarotti. Clock Around the Clock: Time-Based browser fingerprinting. Conference on Computer and Communications Security (CCS '18). <https://dl.acm.org/citation.cfm?id=3243796>
- [16] V. Mishra, P. Laperdrix, A. Vastel, W. Rudametkin, R. Rouvoy, M. Lopatka.. Don't count me out: On the relevance of IP addresses in the tracking ecosystem. The Web Conference 2020 (WWW'20). <https://hal.inria.fr/hal-02435622>
- [17] The Privacy Sandbox Timeline for the Web. 2023-11 (accessed 2023-11-30). https://privacysandbox.com/intl/en_us/open-web/#the-privacy-sandbox-timeline
- [18] L. Von Ahn, M. Blum, N. J. Hopper, and J. Langford, "Captcha: Using hard ai problems for security," in International Conference on the Theory and Applications of Cryptographic Techniques. Springer, 2003, pp. 294–311
- [19] G. Wang, T. Konolige, C. Wilson, X. Wang, H. Zheng, and B. Y. Zhao, "You are how you click: Clickstream analysis for sybil detection." in USENIX Security Symposium, vol. 9, 2013, pp. 1–008.
- [20] A. Stassopoulou and M. D. Dikaiakos, "Web robot detection: A probabilistic reasoning approach," Computer Networks, vol. 53, no. 3, pp. 265–278, 2009
- [21] D. Stevanovic, A. An, and N. Vljajic, "Feature evaluation for web crawler detection with data mining techniques," Expert Systems with Applications, vol. 39, no. 10, pp. 8707–8717, 2012.
- [22] A. Balla, A. Stassopoulou, and M. D. Dikaiakos, "Real-time web crawler detection," in Telecommunications (ICT), 2011 18th International Conference on. IEEE, 2011, pp. 428–432.
- [23] Conversion API. (accessed 2023-12-04). <https://developers.facebook.com/docs/marketing-api/conversions-api/>
- [24] Best practices for Conversions API. (accessed 2023-12-04). <https://www.facebook.com/business/help/308855623839366?id=818859032317965>
- [25] About cookie settings for the Meta pixel. (accessed 2023-12-04). <https://www.facebook.com/business/help/471978536642445?id=1205376682832142>
- [26] An introduction to server-side tagging. (accessed 2023-12-04). <https://developers.google.com/tag-platform/tag-manager/server-side/intro>
- [27] Am I Unique ? (accessed 2023-12-04). <https://amiunique.org>

Walter Rudametkin is a Professor in Computer Science at the University of Rennes, France.

Olivier Zendra is a Tenured Computer Science Researcher at Inria, Rennes, France.



Distributed Ledger Technologies (DLT) and InterPlanetary File System (IPFS) are key enablers to achieve an overall decentralized, secured platform over the Internet, like the Next computing Paradigm (NCP).

DLT and IPFS Technologies are Paving the Way for the Next Paradigm in Computing

by Emmanuel Bertin

Key insights

- Distributed Ledger Technologies (DLT) and InterPlanetary File System (IPFS) are key enablers to achieve an overall decentralized, secured platform over the Internet.
- Mass adoption of this new computing paradigm will require thorough transformation of European laws and policies, including on the legal value of smart contracts.
- Many challenges are still ahead to pave the road to this new computing paradigm.

Key recommendations

- DLT and IPFS allow combining security and controlled openness. EU companies should consider these technologies as a driver to rethink their business models, relying on mission-oriented consortiums rather than on mere competition, and moving from “paper-based partnerships” (e.g., through contracts and agreements) to “code-base partnership” (e.g., using smart-contracts).
- Similarly, EU public authorities (including regulation bodies) should consider moving from “paper-based policies” to “code-base policies” (e.g., with smart-contracts) relying on DLT and IPFS. They may also consider becoming part of DLT consortiums, in order to monitor the dynamic enforcement of policies.
- Investigate and monitor possible ways, especially security issues, to circumvent European policies using Blockchain and IPFS technologies, whether by European or foreign actors

Key components for a decentralized internet paradigm

Distributed Ledger Technologies (DLT) and InterPlanetary File System (IPFS) are not just technologies; they are the building blocks of a new computing paradigm that promises to reshape our digital landscape. These technologies are indeed at the forefront of building a distributed computer over the entire internet. This concept, often referred to as the "world computer" envisions a future where decentralized applications (dApps) run on a peer-to-peer network, leveraging the collective processing power and storage capacity of millions of devices connected to the internet [1].

DLT refers to a digital system for recording the transaction of assets in which the transactions and their details are recorded in a decentralized manner, each peer holding a synchronized copy of every transaction. It enables the creation of a distributed ledger that records data across a network of nodes, ensuring that no single point of failure can compromise the integrity of the data. Unlike traditional databases, DLT has no central data store or administration functionality. Blockchain, which underpins cryptocurrencies like Bitcoin, is one of the most well-known types of DLT.

IPFS complements blockchain by addressing one of its limitations: storage. While blockchain is excellent for recording transactions and small amounts of data, it is not efficient for storing large files. IPFS is a protocol and network designed to create a content-addressable (meaning files are retrieved based on their content address rather than their location), peer-to-peer method of storing and sharing hypermedia in a distributed file system. IPFS enables files to be stored across multiple nodes globally, which can reduce redundancy, increase efficiency, and improve the web's permanence.

The "World Computer" Concept: The idea of a "world computer" stems from the vision of a decentralized computing infrastructure. In this paradigm, dApps (decentralized applications) operate on a global peer-to-peer network, utilizing the combined processing power and storage of millions of devices. This can potentially offer benefits such as increased resilience against failures, cen-

sorship resistance, and a new model for digital interactions and transactions without the need for traditional centralized intermediaries. This vision could be a way to achieve the Next Computer Paradigm (NCP). While the NCP is focused on a global nest of services, systems and devices, this idea of "world computer" relying on DLT and IPFS can be seen as one of the enablers to implement the NCP vision (e.g., for achieving global and decentralized computing and storage, as well as applications built on top of them).

Together, DLT and IPFS are foundational to enable this vision by respectively providing mechanisms for secure, decentralized transaction recording, and file storage and distribution. However, governance issues have to be considered.

Governance in a Distributed World

In a distributed network, governance refers to the mechanisms through which decisions are made, changes are implemented, and conflicts are resolved. The challenges include ensuring that the system is fair, transparent, and efficient while preventing any single entity from gaining too much control.

Various models have emerged to address these challenges:

- **On-chain governance:** Decisions are made through a formalized process coded into the blockchain itself, often involving token-based voting¹. When disagreements arise, the blockchain may undergo a "fork," where the ledger diverges into two separate chains, each reflecting different decisions or rule sets.
- **Off-chain governance:** Decisions are made through discussions and agreements among the community in charge of the network, and then implemented by the network's validators or developers.

These different governance models can be illustrated with some famous public blockchain projects, featuring different governance structures:

¹ Such on-chain governance is the ground for DAO (Distributed Autonomous Organizations) [5], where this token-based voting model is applied to the governance of a whole organization. However, this

type of organization has not yet proven a widescale and sustainable viability.

- Bitcoin [2]: Utilizes an informal off-chain governance model where the community, developers, and miners discuss and implement changes through consensus, with the possibility of forks.
- Ethereum [3]: Has a mix of on-chain and off-chain governance, with formal improvement proposals and community discussions leading to upgrades like Ethereum 2.0.
- Tezos [4]: Features an on-chain governance model where token holders vote on proposals for protocol upgrades that are then automatically implemented without forking.

In addition, DLT may also be deployed and operated in a private way by a consortium of actors (typically a set of companies). Those cases mainly rely on a pure off-chain governance, to let the involved companies shape the consortium. However, the detailed governance models are not publicly documented and remain specific to each consortium.

We believe here there is an opportunity for EU public authorities and companies to work together on this topic of governance. On-chain governance appears clearly as a very promising scenario, enabling to automate many back-office tasks. However, its limitations and its lawful status should be addressed in the EU scope. The possibility to include in Blockchain consortiums a national or EU public authority, which would be in charge to monitor the governance model, and enforce it in litigation cases, should also be considered.

Sovereignty and Data Ownership

DLT and IPFS can empower users with sovereignty over their data by enabling them to control where and how their data is stored and shared.

- **DLT:** Provides a secure and immutable ledger where users can prove ownership of data or assets without the need for a central authority. Smart-contracts can enforce the rules of data access and usage.
- **IPFS:** Allows users to store data on a distributed network where they control who has access to it. Data is addressed by content rather than location, making it resilient to censorship and server failure. Decentralization enables a high scalability and a low latency for reading accesses (writing accesses are immediate, but

there is of course a synchronization delay, as in any peer-to-peer system).

The implications for privacy and data security include:

- **Enhanced Privacy:** Users can control their personal data and share it selectively, potentially reducing the risk of mass data collection by centralized entities.
- **Improved Data Security:** A distributed network reduces the risk of centralized data breaches. Data is spread across multiple nodes, making it harder to compromise.

However, these technologies also introduce new challenges, such as the need for robust encryption and secure key management, as users are responsible for their own data security.

Theoretical and Operational Challenges

More precisely, the challenges associated with enhanced privacy and improved data security in a distributed network [6] like those enabled by DLT and IPFS include the following ones.

Secure Key Management: Users are responsible for their own private keys, which are required to access their data and assets. Losing a private key can result in the permanent loss of access to data or digital assets. Storing keys securely while making them accessible and recoverable is a complex issue.

Data Availability and Persistence: In a distributed system, data is replicated among multiple nodes to ensure it remains accessible even if some nodes go offline. Ensuring that data effectively persists on the network over time remains a long-term challenge, especially on DLT that are not massively used.

Scalability: As the number of transactions and users grows, maintaining performance and efficiency is a challenge. Blockchain networks, in particular, can suffer from slow transaction times and higher fees during periods of congestion. For example, the large-scale public Bitcoin network can proceed about 7 transactions per second, while Visa processes around 1,700 transactions per second on average, claiming to be able to support 24,000. However, this issue can be partly circumvented by relying on layer 2 DLT, that are built as

an overlay on top of an existing Blockchain network.

Regulatory Compliance: Data sovereignty laws and regulations, such as GDPR, may impose requirements on data storage and transfer that are difficult to reconcile with decentralized networks. Ensuring compliance while maintaining the decentralized ethos poses a significant challenge. In particular, the EU “right to be forgotten” (RTBF) is not easily compatible with the data immutability property of DLT (meaning that once written in the chain, a data can never be deleted).

User Experience: The complexity of managing one's own data and security can be a barrier to adoption for less technical users. Developing user-friendly interfaces and processes that do not compromise security is crucial for wider acceptance. Key management being here clearly an issue to achieve a great ease of use while maintaining a high security level.

Smart-Contract Security: Smart-contracts are immutable once deployed to a blockchain, and bugs or vulnerabilities can lead to loss of funds or data breaches. Ensuring the security of smart-contracts requires rigorous testing and auditing before deployment.

Network Security: Decentralized networks must be resilient to various attacks, such as 51% attacks [7] (meaning that more than 51% of the nodes are malicious), Sybil attacks [8] (meaning that a single malicious entity controls directly or indirectly a large number of nodes), and routing attacks [9] (meaning attacks at the IP routing level in order to block communication between nodes and split the network). While these attacks are well documented, and addressed in current DLT and IPFS networks, maintaining a secure and robust network requires constant vigilance and adaptation to new threats.

Addressing these challenges is an ongoing process that involves both technological innovation and changes in user behavior. As these technologies mature, solutions to these challenges are being developed, but the responsibility for data security and privacy increasingly falls on the individual user. For example, as DLT are massively relying on asymmetric cryptography (e.g., any operation must be signed by user's private key), the loss of his or her private key by the user of a public blockchain implies the loss of any ability to act on the chain (including retrieving funds); systems that

could help to circumvent the problem are often hardly compatible with the decentralization principles of DLT).

Early Use Cases

DLT and IPFS are being utilized across various industries, each with specific applications and business benefits.

In Finance, Blockchain is used for cryptocurrencies, cross-border payments, and smart-contracts that automate complex financial transactions [10]. It reduces transaction times and costs while increasing transparency. In Supply Chain, Blockchain provides traceability and transparency from production to delivery, helping to verify the authenticity of products and streamline logistics [11]. For Healthcare, IPFS can store large medical files in a decentralized manner [12]. Concerning Real Estate, Blockchain may simplify property transactions by reducing the need for intermediaries, automating land registry updates, and ensuring the immutability of records [13]. For Media and Entertainment, IPFS helps content creators distribute their work directly to consumers, potentially reducing hosting costs and improving content addressability [14].

However, these use cases are not so easily converted into business. For example, Maersk and IBM's TradeLens platform uses blockchain to streamline maritime shipping was stopped in November 2022, nearly five years after the creation of a dedicated joint-venture to create and deploy this platform. The reason invoked for this closure was not technical, but about the lack of involvement of other business partners, leading to a poor commercial viability: “TradeLens was founded on the bold vision to make a leap in global supply chain digitization as an open and neutral industry platform. Unfortunately, while we successfully developed a viable platform, the need for full global industry collaboration has not been achieved. As a result, TradeLens has not reached the level of commercial viability necessary to continue work and meet the financial expectations as an independent business.” [15]

Conclusion

The combination of blockchain and IPFS creates a synergistic effect that can revolutionize how we think about computing resources. Instead of relying on centralized servers and data centers, a distributed computer over the internet would harness the unused computational resources and storage space of participating devices, creating a more resilient and efficient system.

This distributed approach also has profound implications for governance and sovereignty. In a world where data and applications are not tied to any single entity, users regain control over their digital assets. This shift challenges traditional models of operation and requires new frameworks for governance that can accommodate the decentralized nature of these technologies.

However, the path to realizing this vision is not without obstacles. Scalability, interoperability, and user adoption remain significant challenges. Blockchain networks, in particular, must find ways to handle an increasing number of transactions without compromising speed or security. Similarly, IPFS must ensure that data retrieval remains fast and reliable as the network grows.

In addition, mass adoption of this new computing paradigm will require thorough transformation of European laws and policies, including on the legal value of smart-contracts, and also to detail how existing norms can be applied, including the “right to be forgotten”. Possible ways to circumvent European policies using these technologies, whether by European or foreign actors, should however be carefully investigated and monitored, especially concerning security issues.

Concerning Public authorities (including regulation bodies, in particular for sectorial regulation), we recommend they progressively consider moving from “paper-based policies” to “code-base policies”, by framing consortiums of business actors, relying on Blockchain and IPFS technologies. For example, Mobile Number Portability could be performed by smart-contracts with encrypted data stored on IPFS, with the underlying Blockchain being implemented as a consortium among Mobile Network Operators and regulatory bodies for each country, as proposed in [16].

More broadly, EU companies should consider Blockchain and IPFS as an opportunity to rethink their models, relying more on mission-oriented

consortiums rather than on mere competition, as these technologies are now enabling to couple security, control, and (more or less) openness. Traceability, recycling or logistics are here typical examples.

References

- [1] Julien Hatin, Emmanuel Bertin, Baptiste Hemery, Nour El Madhoun. Welcome to the jungle: A Reference Model for Blockchain, DLT and Smart-Contracts. Tokenomics 2020 on Blockchain Economics, Security & Protocols (2nd International Conference), Oct 2020, Toulouse, France. <https://hal.science/hal-03014631>
- [2] What is Bitcoin governance? <https://www.bitcoin.com/get-started/what-is-bitcoin-governance/>
- [3] Introduction to Ethereum governance. <https://ethereum.org/en/governance/>
- [4] Governance and validation on Tezos. <https://tezos.com/governance/>
- [5] Hassan, Samer, and Primavera De Filippi. 2021. "Decentralized Autonomous Organization". *Internet Policy Review* 10 (2). DOI: 10.14763/2021.2.1556. <https://policyreview.info/glossary/DAO>
- [6] Huang, Huawei, Jianru Lin, Baichuan Zheng, Zibin Zheng, and Jing Bian. "When blockchain meets distributed file systems: An overview, challenges, and open issues." *IEEE Access* 8 (2020): 50574-50586. <https://ieeexplore.ieee.org/document/9031420>
- [7] Sayeed, Sarwar, and Hector Marco-Gisbert. "Assessing blockchain consensus and security mechanisms against the 51% attack." *Applied sciences* 9, no. 9 (2019): 1788.
- [8] Iqbal, Mubashar, and Raimundas Matulevičius. "Exploring sybil and double-spending risks in blockchain systems." *IEEE Access* 9 (2021): 76153-76177.
- [9] Chaganti, Rajasekhar, Rajendra V. Boppana, Vinayakumar Ravi, Kashif Munir, Mubarak Almutairi, Furqan Rustam, Ernesto Lee, and Imran Ashraf. "A comprehensive review of denial of service attacks in blockchain ecosystem and open challenges." *IEEE Access* 10 (2022): 96538-96555.
- [10] Patel, Ritesh, Milena Migliavacca, and Marco E. Oriani. "Blockchain in banking and finance: A bibliometric review." *Research in International Business and Finance* 62 (2022): 101718.
- [11] Raja Santhi, Abirami, and Padmakumar Muthuswamy. "Influence of blockchain technology in manufacturing supply chain and logistics." *Logistics* 6, no. 1 (2022): 15.
- [12] Jayabalan, Jayapriya, and N. Jeyanthi. "Scalable blockchain model using off-chain IPFS storage for healthcare data security and privacy." *Journal of Parallel and Distributed Computing* 164 (2022): 152-1
- [13] Garcia-Teruel, Rosa M. "Legal challenges and opportunities of blockchain technology in the real estate sector." *Journal of Property, Planning and Environmental Law* 12, no. 2 (2020): 129-145.
- [14] Dutra, Andre, Andranik Tumasjan, and Isabell M. Welpe. "Blockchain is changing how media and entertainment companies compete." *MIT Sloan Management Review* (2018).
- [15] Maersk and IBM to discontinue TradeLens, a blockchain-enabled global trade platform. <https://www.maersk.com/news/articles/2022/11/29/maersk-and-ibm-to-discontinue-tradelens>
- [16] Ghaffari, Fariba, Emmanuel Bertin, and Noel Crespi. "User Profile and Mobile Number Portability for Beyond 5G: Blockchain-based Solution." In *2023 26th Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN)*, pp. 187-194. IEEE, 2023.

Emmanuel Bertin is Senior Expert at Orange, Paris, France, as well as associated Professor at Telcom Sud Paris, Palaiseau, France.



Hardware constitutes the foundation of any computer system. Ensuring its integrity throughout the entirety of the hardware supply chain poses a significant challenge in establishing a secure computer system. The involvement of numerous untrusted parties in the process opens the door to vulnerabilities.

Integrity at Every Link: A Roadmap to Trustworthy Hardware Supply Chains

by Lennart M. Reimann, Dominik Sisejkovic and Rainer Leupers

The Integrated Circuits (ICs) supply chain uses a horizontal model, where Intellectual Property (IP) owners rely on external partners for competitiveness and cost reduction. However, this reliance raises significant trust concerns, including IP theft, IC counterfeiting, and the introduction of malicious circuit alterations (Hardware Trojans (HTs)) [1].

In the last decade, Hardware Trojans have emerged as a significant security concern [1][2]. These circuit alterations pose a threat by allowing unauthorized access, manipulation, and control of electronic systems. While practical instances of hardware Trojans are not conclusively documented, the ability to make subtle modifications with basic tools [3] emphasizes the need for robust security measures.

Over the past decades, numerous research programs have explored different methodologies aimed at ensuring trustworthiness throughout the IC supply chain. For instance, the United States Defense Advanced Research Project Agency (DARPA) has initiated multiple funding programs to advance R&D in the domain of reliable electronics. These programs include IRIS [4], TRUST [5], and SHIELD [6], among others. The significance of this issue has also been acknowledged in Germany, with the German Federal Ministry of

Education and Research (BMBWF) launching a funding program spanning 2021 to 2024, specifically aimed at addressing the challenges of dependable microelectronics for Germany and Europe [7] [8]. Regrettably, as of now, there is no established formal process for ensuring the trustworthiness of hardware across the IC supply chain. Consequently, it is required to assess existing protective measures and determine the necessary focus areas for future research, with the ultimate goal of guaranteeing the security of hardware throughout the supply chain. This article is based on the findings provided in [11].

Key Insights

- Design-dependent hardware Trojans are a fundamental security issue.
- Standard detection mechanisms only allow the identification of known hardware Trojans.
- Newer approaches focus on identifying malicious modifications by comparing the IP with a golden reference throughout the complete supply chain.
- Automated and complete reverse engineering is crucial for achieving formal security guarantees for the entire supply chain.

- Best-effort security measures implemented through active protection mechanisms remain an important pillar in protecting against malicious hardware modifications.

Key Recommendations

To holistically protect against (1) malicious hardware modifications (Trojans), we need to invest in (2) long-term research to formally secure the hardware supply chain and (3) short-term, best-effort security.

- (1) Invest in a tangible estimation of the effort, tools, and skills required to design and insert design-specific hardware Trojans.
- (2) Formal security guarantees:
 - Develop end-to-end, automatic, zero-fault, and non-destructive reverse engineering methods as key enabler for end-to-end equivalence checking.
 - Develop end-to-end and complete equivalence checking (EQ) methods from the abstract design specifications to the final physical device.
- (3) Best-effort active protection mechanisms: advance the research on best-effort security with active protection mechanisms that aim at disallowing malicious modifications to the design throughout the hardware supply chain.

From Specification to Silicon

The hardware IP necessitates protection across the entire supply chain, spanning from initial specifications to the final device. The supply chain involves the IP owner, design house, foundry, assembly facility, OEMs, and users. The asset undergoes format changes via automated EDA tools or manual processes by hardware designers [9][10].

1. IP Creation and Ownership

- The process starts with the creation of hardware IP, which includes specifications, virtual prototypes, high-level descriptions, Register Transfer Level (RTL) designs, and gate-level netlists.

- This IP is initially owned by a development team, company, or individual who holds the rights to its use and distribution.

2. External Design House

- The IP owner may collaborate with an external design house for specialized expertise or additional resources.
- The IP owner provides either the RTL design or gate-level netlist to the external design house.

3. Design and Layout

- The external design house takes the provided RTL design or gate-level netlist and utilizes it to generate the layout in GDSII format.
- This layout serves as a blueprint for the physical components of the hardware.

4. Foundry and Manufacturing

- The GDSII layout is forwarded to a foundry, which is a specialized facility equipped for semiconductor fabrication.
- The foundry utilizes the layout to develop a mask of the chip design.
- The mask is used to manufacture the final chip using advanced semiconductor fabrication processes.

5. Assembly and Integration

- Once the chips have been produced, they are sent to an assembly facility.
- At this stage, the chips are combined with other electronic components to create either intermediate devices or the final product.

6. Packaging and Distribution

- The assembled devices undergo packaging to protect them from environmental factors and facilitate handling during transportation and use.
- Packaged devices are then distributed to original equipment manufacturers or directly to end users.

7. OEMs

- OEMs may further incorporate the hardware into larger systems or products.

8. End Users

- End users receive the final packaged devices, which they utilize for their intended purposes.

Throughout this journey, the original hardware IP undergoes a series of transformations from high-level descriptions to tangible physical components. It is crucial to ensure the integrity and security of the IP at each stage, as it may be exposed to vulnerabilities when handled by external parties (see Figure 1).

Hardware Trojans: A Fundamental Threat in the Hardware Supply Chain

The electronics supply chain, geared towards minimizing time-to-market and cutting costs, involves various external entities and closed-source third-party Electronic Design Automation (EDA) tools. This decentralized structure introduces security vulnerabilities due to inherent trust uncertainties among involved parties. A significant consequence is the potential introduction of malicious alterations known as hardware Trojans (HTs), an ongoing concern in security research for over a decade [2]. Unfortunately, the issue of HTs remains largely unresolved.

Hardware Trojans: A Subtle Change with Disastrous Consequences

Over ten years ago, a radar system in Syria failed to provide advance warning of an approaching air-strike, purportedly due to the presence of HTs in the defense systems [46]. While confirming the presence of HTs in such instances is challenging, the mere possibility of these subtle yet malicious design alterations has garnered significant attention in both research and industry. The US military and intelligence executives have identified HTs as one of the most significant threats the nation could encounter during times of war.

An HT is characterized by intentional, malicious, and covert alterations made to integrated circuits throughout the entire hardware supply chain [12]. This malicious behavior can take different forms, including information leakage, power dissipation, denial of service, performance degradation, or unintended behaviors. The intentional nature of the modification distinguishes it from random faults. Trojans are implemented stealthily to evade detection during tests and security checks.

Before we take a look at different classes of HTs, it is important to understand the process of Reverse Engineering (RE) - a term that is tightly coupled to HTs. Hardware RE is defined as the process of extracting a set of specifications for a hardware design by an entity other than the original design owner [42]. Consequently, RE has traditionally been associated with potentially malicious activities, such as IP theft, that are of significant

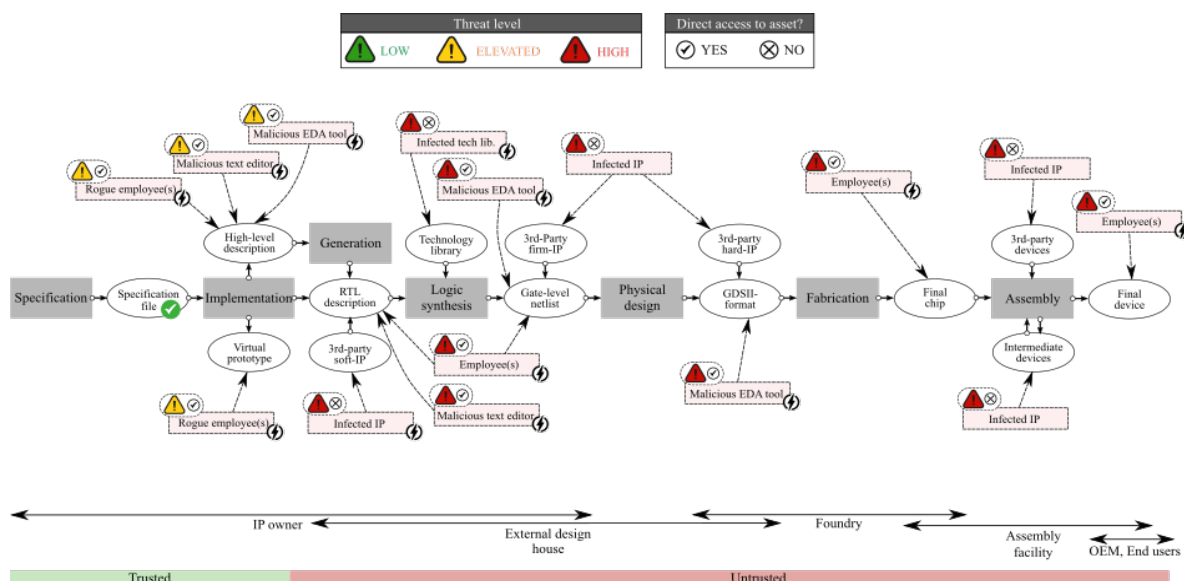


Figure 1 Assets and vulnerabilities within the hardware supply chain [11]

concern to governments, the military, and industry.

Within the context of RE, HTs can be categorized by addressing a fundamental question: does the design and insertion of the Trojan necessitate RE? This classification framework, as outlined in [14], divides all Trojans in two groups: Class-1 HTs (C1HTs) and Class-2 HTs (C2HTs). C1HTs encompass Trojans that rely on RE. As a result, the attacker needs to invest into comprehending the design specifications of the asset at various levels to create an HT tailored to that specific design. Consequently, a Class-1 HT enables a controllable trigger, paving the way for a high-impact attack. C2HTs consist of less hidden Trojans that do not rely on RE. Consequently, an attacker can insert these HTs into a design at any stage or level of abstraction without possessing any knowledge about the asset. Thus, C2HTs remain resilient against potential protection mechanisms.

The Untrusted Design House and Foundry

One of the most common threat scenarios within the hardware supply chain involves the malicious actions of external design houses and foundries [1] [13] [40]. Since these third-party entities are often located at remote sites around the world and lack a verifiable level of trust in the design and production process, they present a potential vulnerability for malicious alterations, such as HTs. Both external design houses and foundries receive the asset in a form that remains modifiable before it is permanently coded into silicon. What are our assumptions about the adversaries' capabilities? In general, the following is true for both an external design house and foundry: (1) The entity is granted complete access to the design. The external design house is provided with either the RTL or gate-level design, while the foundry receives the final layout. (2) The entity functions without any limitations or oversight by the legitimate IP owner. (3) The insertion of a design-specific (class 1) HT demands a certain level of RE effort.

A crucial element influencing the characteristics of the introduced Trojan is RE. The primary aim of RE is to attain an abstraction level of the asset that allows for subsequent analysis and manipulation. As the asset's abstraction level decreases, the need for a more extensive effort becomes apparent to gain a comprehensive understanding of the design and, potentially, to insert class-1 HTs. The RE process contains a multitude of manual, semi-automatic, and automatic steps which all paint only

part of a still error-prone, lengthy picture. Thus, a completely automated, non-destructive, and flawless RE process still remains elusive despite being the key factor in determining the effectiveness of an active protection against HTs.

Open Challenges: Reverse Engineering

- How to quantify the cost, required effort, and complexity of reverse engineering?
- How to quantify the success criteria and the amount of retrieved information of the reverse engineering process?
- How to implement non-destructive, automatic, and zero-fault RE for every abstraction level of the asset?

Fighting Hardware Trojans

In the pursuit of combating HTs, research takes two distinct approaches. While some design methodologies focus on detecting these insidious elements for subsequent removal, achieving fool-proof detection faces formidable challenges along the hardware supply chain. Consequently, many researchers are now delving into proactive measures aimed at safeguarding against malicious modifications in the first place.

Detecting Hardware Trojans

A high number of methodologies aim to prevent HTs by identifying malicious modifications in the asset and removing them. Detection mechanisms fall into two classes: pre-silicon and post-silicon. Pre-silicon methods focus on design analysis before manufacturing, while post-silicon mechanisms target the manufactured design.

Pre-Silicon Detection

Pre-silicon detection mechanisms can be summarized with the following four major techniques: Code coverage analysis, formal verification techniques, structural analysis, and functional analysis.

In this context, formal verification stands out as the most promising approach for thoroughly proving the absence of hardware Trojans. Formal verification techniques like equivalence checking can mathematically prove properties and equivalence between two descriptions of the asset. This allows detecting discrepancies indicating potential

Trojans. Existing verification methods like security assertions can also be reused for Trojan detection [17] [18] [45]. However, limited abstraction details and a lack of formal models for manufactured chips constrain formal techniques. A common formal method used in this context is equivalence checking. Equivalence checking in hardware security verifies that two different representations of a design perform the same operations and produce the same results [15] [16]. It ensures the integrity and trustworthiness of a design, identifying any discrepancies or potential malicious modifications. Therefore, it allows comparing two descriptions of the same IP between any two steps in the supply chain.

Moreover, approaches like code coverage analysis, structural analysis, and functional tests are employed to identify hardware Trojans [19][20]. However, due to their limitations in providing comprehensive security assurance or necessitating specific Trojan structure information, there is an increasing research emphasis on formal methods.

Post-Silicon Detection

Functional tests execute application tests on the chip to check for incorrect behaviors indicating potential Trojans. But tests may not trigger all Trojans, and some may not corrupt functionality.

After the chips are manufactured every device needs to be tested for modifications instead of analyzing a single hardware description. Therefore, researchers work on reverse-engineering to yield a formal description of the produced hardware. Reverse engineering involves delayering and imaging the manufactured chip layer-by-layer to reconstruct a gate-level netlist [1] and apply pre-silicon analysis. However, RE is expensive, time-consuming, and destructive.

Overall, post-silicon detection has limited coverage compared to pre-silicon techniques. RE provides a gate-level netlist but is expensive and destructive. More research is needed to enable post-silicon security guarantees.

Open Challenges: Passive Detection Mechanisms

- Does equivalence checking offer a complete assurance of the absence of malicious modifications?

- Is there a detection scheme capable of identifying any potential hardware Trojan?

Protection Against Malicious Design Modifications

Many Design-for-Trust (DfTr) methodologies have been introduced in the last decades to protect hardware against malicious modifications, including functional filler cells [37] [39], split manufacturing [40] [41], and layout camouflaging [42] [43]. In the following, we will, however, only take a closer look at logic locking – a premier technique to circumvent the insertion of class 1 HTs, and the only active protection mechanism capable of protecting against untrusted entities throughout the microelectronics supply chain

Logic Locking (LL) aims to protect the integrity of hardware designs at different supply-chain stages and design levels [38]. LL modifies the hardware design through the incorporation of logic alterations that link the proper functioning of the chip to a confidential activation key. This alteration carries two primary consequences. First, the functional behavior of the HW design is contingent upon the correctness of the key. When the correct key is applied, the design operates as intended. Second, the inclusion of key-dependent logic brings about structural modifications in the design, essentially "obfuscating" the hardware.

How is LL applied? Let's assume that logic locking is implemented at the gate level. The IP owner – the trusted entity - aims to develop a legitimate chip. During this stage, the RTL description of the hardware design is logically synthesized into a gate-level netlist. At that point, logic locking is applied, resulting in a locked netlist and a secret key. The secret key remains exclusively with the legitimate IP owner. Note that the key is not needed for any subsequent steps. The locked netlist is then provided to external parties for layout generation, fabrication, and assembly. Once the final chip is prepared, the IP owner performs activation. The secret key is incorporated into the chip using a non-volatile memory, such as flash, e-fuse, or EEPROM [23]. This process has been successfully implemented by HENSOLDT Cyber GmbH through the production of the "Made in Germany RISC-V" (MiG-V) processor—a groundbreaking example of a fully logic-locked commercial processor [14] [24] [25] [32].

The Interplay of Reverse Engineering and Logic Locking

A lot of research has popped up around the topic of logic locking, including resilient schemes design and novel key recovery attacks [44]. Unfortunately, one mistake is still being repeated: the security of LL is only seen through the recoverability of the key. Why is this a problem? The key itself should evidently not be recoverable before production; otherwise, an attacker could simply remove the locking-induced structures and dissolve the impact of the obfuscation. However, focusing on “how difficult is it to retrieve the key” as a measure of security is somewhat misleading as it completely ignores the main objective of LL: making reverse engineering harder to perform. As a result, the concept of the key's "retrievability" has often been employed as an indicator of security. However, the question of "how much more challenging RE becomes" because of logic locking has remained unanswered.

Secret Key, Unsecure Storage

A fundamental issue in logic locking hides in the availability of a secure key storage. Unfortunately, a growing number of physical attacks have successfully shown that the correct key can be extracted from an activated chip through probing and fault-injection attacks [27] [28] [29]. Moreover, it is possible to design a design-independent HT that leaks the key value after the chip is activated simply by forwarding the key inputs to an observable output [30]. Hence, without a secure key storage, logic locking will have a very limited effect in a high-volume production setting which allows the availability of activated chips on the market.

Universal Circuits: A Way Out?

A promising approach to addressing many challenges in logic locking is rooted in the concept of universal circuits [26] [31]. Drawing from a cryptographic primitive introduced by Valiant [33], universal circuits can be programmed to emulate any circuit within a specified size limit. From a security perspective, a universal circuit can represent a wide range of hardware functionalities while consistently maintaining the same underlying structure. In fact, universal circuits could be seen as “the ultimate” obfuscation. Why is this interesting? When all the circuits entrusted to a potentially untrustworthy foundry or external design house share the same physical structure,

irrespective of the functionality programmed by a secret key, an attacker has only one avenue for introducing modifications: random, design-independent, and most likely low-impact class 2 HTs. Regrettably, the expenses associated with the implementation of this approach far exceed acceptable levels. As a compromise, an alternative solution has been examined using Embedded Field-Programmable Gate Arrays (eFPGAs). In this setup, specific security-critical design modules are substituted with fully reconfigurable soft eFPGA or pre-designed eFPGA hard macros [22] [34]. However, additional research is imperative to ascertain the security and cost-effectiveness of FPGA-based obfuscation [22] [35] [36].

Open Challenges: Active Protection Mechanisms

- How to protect the activation key in logic locking from physical attacks?
- How to measure the impact of logic locking on the required reverse-engineering effort?
- How to design cost-effective and generic indistinguishable circuits?

The Silver Bullet: Formal Guarantees

The ultimate goal in security is to achieve formal guarantees for the absence of malicious modifications. Formal verification is the major approach to achieving this goal. However, although a mathematical proof for the analysis is given, there are a few points that are still open for research to achieve this complete guarantee.

1. As the IP changes its level of abstraction throughout the hardware supply chain, the discrepancy between the levels of abstraction needs to be considered.
2. Manufactured chips do not offer a formal description of the underlying hardware so that reverse engineering needs to be advanced further to offer a non-destructible, fast and complete solutions to generate a description for every manufactured device
3. A complete chain of formal verification tools is required to cover the entire hardware supply chain. This needs to be standardized and further development is required.

Research Needs: The Way Forward

To establish a secure microelectronics supply chain, it is essential to address two fundamental challenges. (1) The first challenge revolves around formally securing the entire supply chain and achieving mathematically proven security assurances across the entirety of the hardware design and manufacturing process. Undoubtedly, pursuing this objective embodies the highest level of security, although it may involve high risks and long-term projects. (2) Consequently, it is equally vital to promote research endeavors focusing on lower-risk, short-term projects that aim to provide best-effort security. These two overarching goals are further elucidated below, following the visualization in Figure 2.

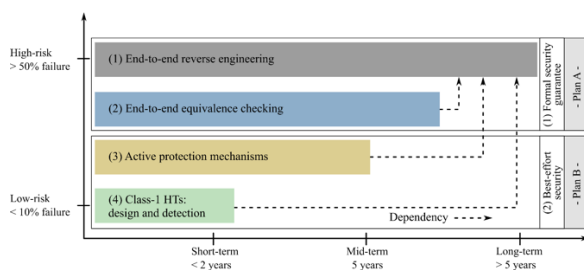


Figure 2 Research goals for achieving a secure hardware supply chain [11].

Reaching Formal Security Guarantees

To enable a formally secure hardware design and fabrication flow, the following must be achieved:

Need 1: A fully automated, error-free, non-destructive, and seamless reverse engineering process spanning from the physical device to high-level abstractions is of paramount importance. The establishment of a comprehensive RE workflow, starting from the final chip, serves as a critical facilitator for comprehensive Equivalence Checking (EQ), HT detection, and the evaluation of active protection approaches. This research gap includes the following tasks:

- Estimation of the complexity, cost, and time effort of reverse engineering;
- Design and implementation of fully automatic reverse engineering methodologies.

Need 2: Achieving end-to-end and comprehensive EQ spanning from high-level abstractions down to the ultimate physical device is imperative.

This EQ continuum guarantees that the final, manufactured, packaged, and embedded device maintains complete equivalence with its initial design specifications. This research gap includes the following tasks:

- Evaluation of the influence of the abstraction level and the design details it provides on the effectiveness of equivalence checking;
- Introducing formats and standards that enable equivalence checking at high abstraction levels;
- Introducing formats, standards, and methods to enable post-fabrication equivalence checking;
- Offering open-source, verifiable, and trustworthy equivalence-checking tools.

Supporting Best-Effort Security

It is crucial to back research endeavors addressing lower-risk gaps that may not lead to formal and all-encompassing security assurances but contribute to best-effort security. These research objectives encompass:

Need 3: Active protection mechanisms may not offer formal security assurances, yet they represent a crucial and currently the sole line of defense against malevolent alterations in the course of external design and manufacturing stages. Nonetheless, the efficiency of active protection methods is tethered to several unresolved inquiries, such as:

- Enabling secure key storage solutions that are resilient against physical attacks;
- Supporting the development of cost-efficient universal circuits or approximations thereof in the form of reconfigurable circuits;
- Evaluate the impact of active protection mechanisms on the reverse engineering effort;
- Evaluate the possibility of formally secure active protection mechanisms.

Need 4: The design and detection of class-1 HTs remains a focal point in security research. To facilitate this goal, the following must be considered:

- A concrete assessment of the resources, tools, and expertise needed for the creation and insertion of Class-1 HTs;

- b. Evaluation of the design of HTs that could potentially circumvent equivalence checking;
- c. Support post-fabrication, non-equivalence-checking-based HT detection methods.

References

- [1] Swarup Bhunia and Mark Tehranipoor. *Hardware Security: A Hands-on Learning Approach*. 1st. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2018.
- [2] Swarup Bhunia and M Tehranipoor. "The Hardware Trojan War: Attacks, Myths, and Defenses". In: Springer International Publishing (2018). doi: 10.1007/978-3-319-68511-3.
- [3] Tiago Perez and Samuel Pagliarini. "Hardware Trojan Insertion in Finalized Layouts: a Silicon Demonstration". In: arXiv, 2021. doi: 10.48550/ARXIV.2112.02972
- [4] Defense Advanced Research Project Agency (DARP). Integrity and Reliability of Integrated Circuits (IRIS). <https://www.darpa.mil/program/integrity-and-reliability-of-integrated-circuits>. accessed: July 2021.
- [5] Defense Advanced Research Project Agency (DARP). Trusted Integrated Circuits (TRUST). <https://www.darpa.mil/program/trusted-integrated-circuits>. accessed: July 2021
- [6] Defense Advanced Research Project Agency (DARP). Supply Chain Hardware Integrity for Electronics Defense (SHIELD). <https://www.darpa.mil/program/supply-chainhardware-integrity-for-electronics-defense>. accessed: July 2021.
- [7] Bundesministerium für Bildung und Forschung (BMBF). Mikroelektronik. Vertrauenswürdig und nachhaltig. Für Deutschland und Europa. Rahmenprogramm der Bundesregierung für Forschung und Innovation 2021-2024. <https://www.elektronikforschung.de/rahmenprogramm>. accessed: July 2021.
- [8] Bundesministerium für Bildung und Forschung (BMBF). Vertrauenswürdige Elektronik. Forschung und Innovation für technologische Souveränität. <https://www.elektronikforschung.de/service/publikationen/vertrauenswuerdige-elektronik>. accessed: July 2021.
- [9] Bicky Shakya et al. "Benchmarking of hardware trojans and maliciously affected circuits". In: *Journal of Hardware and Systems Security* 1.1 (2017), pp. 85–102. doi: 10.1007/s41635-017-0001-6.
- [10] Imran Abbasi et al. "TrojanZero: Switching Activity-Aware Design of Undetectable Hardware Trojans with Zero Power and Area Footprint". In: *Mar.* 2019, pp. 914–919. doi: 10.23919/DATE.2019.8714829.
- [11] Dominik Sisejkovic et al. *Agentur für Innovation in der Cybersicherheit GmbH. Ecosystem for Trustworthy IT; Los 4: Formal Security Guarantees for Trustworthy Hardware Supply Chains*. <https://www.cyberagentur.de/wp-content/uploads/2023/07/OevIT-Vorstudien-Los-4.pdf>
- [12] K. Xiao et al. "Hardware Trojans: Lessons Learned after One Decade of Research". In: *ACM Trans. Des. Autom. Electron. Syst.* 22.1 (May 2016). doi: 10.1145/2906147.
- [13] Jan-Peter Kleinhans and Nurzat Baisakova. *The global semiconductor value chain*. https://www.stiftung-nv.de/sites/default/files/the_global_semiconductor_value_chain.pdf. Accessed: Oct. 2020.
- [14] Dominik Sisejkovic and Rainer Leupers. *Logic Locking: A Practical Approach to Secure Hardware*. In: Springer Cham (2022). doi: [https://doi.org/10.1007/978-3-031-19123-7\[42\]](https://doi.org/10.1007/978-3-031-19123-7[42])
- M. G. Rekoff. "On reverse engineering". In: *IEEE Transactions on Systems, Man, and Cybernetics SMC-15.2* (1985), pp. 244–252. doi: 10.1109/TSMC.1985.6313354.
- [15] Synopsys Inc. *Formality*. <https://www.synopsys.com/implementation-and-signoff/signoff/formality-equivalence-checking.html>. Oct. 2022.
- [16] Claire Wolf. *Yosys*. <http://bygone.clairixen.net/yosys/documentation.html>. Oct. 2022.
- [17] Lennart M. Reimann et al. "QFlow: Quantitative Information Flow for Security-Aware Hardware Design in Verilog". In: *2021 IEEE 39th International Conference on Computer Design (ICCD)*. 2021, pp. 603–607. doi: 10.1109/ICCD53106.2021.00097
- [18] L. M. Reimann, S. Erdönmez, D. Sisejkovic and R. Leupers, "Quantitative Information Flow for Hardware: Advancing the Attack Landscape," *2023 IEEE 14th Latin America Symposium on Circuits and Systems (LASCAS)*, Quito, Ecuador, 2023, pp. 1-4, doi: 10.1109/LASCAS56464.2023.10108235.
- [19] Huili Chen et al. *AdaTest: Reinforcement Learning and Adaptive Sampling for On-chip Hardware Trojan Detection*. 2022. doi: 10.48550/ARXIV.2204.06117.
- [20] Susmit Jha and Sumit Kumar Jha. "Randomization Based Probabilistic Approach to Detect Trojan Circuits". In: *2008 11th IEEE High Assurance Systems Engineering Symposium*. 2008, pp. 117–124. doi: 10.1109/HASE.2008.37.
- [21] Synopsys Inc. *TestMAX*. <https://www.synopsys.com/implementation-and-signoff/test-automation/testmax-atpg.html>. Oct. 2022.
- [22] Hadi Mardani Kamali et al. "Advances in Logic Locking: Past, Present, and Prospects". In: *Cryptology ePrint Archive* (2022)
- [23] M. Tanjidur Rahman et al. "Defense-in-Depth: A Recipe for Logic Locking to Prevail". In: *Integr. VLSI J.* 72.C (May 2020), pp. 39–57. doi: 10.1016/j.vlsi.2019.12.007.
- [24] HENSOLDT Cyber GmbH. *Press Release: HENSOLDT Cyber presents MiG-V, the first RISC-V Processor "Made in Germany" for Security Applications*. <https://hensoldt-cyber.com/wp-content/uploads/2020/05/20200515-HENSOLDT-Cyber-PM-MiG-V-is-ready-1.pdf>. Accessed: May 2020.
- [25] Dominik Sisejkovic et al. "A Secure Hardware-Software Solution Based on RISC-V, Logic Locking and Microkernel". In: *Proceedings of the 23th International Workshop on Software and Compilers for Embedded Systems. SCOPES '20*. St. Goar, Germany: Association for Computing Machinery, 2020, pp. 62–65. doi: 10.1145/3378678.3391886.
- [26] Kaveh Shamsi, David Z. Pan, and Yier Jin. "On the Impossibility of Approximation Resilient Circuit Locking". In: *2019 IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*. 2019, pp. 161–170. doi: 10.1109/HST.2019.8741035.
- [27] Susanne Engels, Max Hoffmann, and Christof Paar. *The End of Logic Locking? A Critical View on the Security of Logic Locking*. *Cryptology ePrint Archive*, Report 2019/796. <https://eprint.iacr.org/2019/796>. 2019.
- [28] M. T. Rahman et al. "The Key is Left under the Mat: On the Inappropriate Security Assumption of Logic Locking Schemes". In: *2020 IEEE HOST*. 2020, pp. 262–272. doi: 10.1109/HOST45689.2020.9300258.
- [29] A. Jain, M. T. Rahman, and U. Guin. "ATPG-Guided Fault Injection Attacks on Logic Locking". In: *2020 IEEE Physical Assurance and Inspection of Electronics (PAINE)*. 2020, pp. 1–6. doi: 10.1109/PAINE49178.2020.9337734.

- [30] Ayush Jain, Ziqi Zhou, and Ujjwal Guin. "TAAL: Tampering Attack on Any Key-Based Logic Locked Circuits". In: 26.4 (Mar. 2021). doi: 10.1145/3442379.
- [31] Elisaweta Masserova et al. "Logic Locking-Connecting Theory and Practice". In: Cryptology ePrint Archive (2022).
- [32] Šišejković, D., Merchant, F., Reimann, L.M., Leupers, R., Kegreiß, S. (2020). Scaling Logic Locking Schemes to Multi-module Hardware Designs. In: Brinkmann, A., Karl, W., Lankes, S., Tomforde, S., Pionteck, T., Trinitis, C. (eds) Architecture of Computing Systems – ARCS 2020. ARCS 2020. Lecture Notes in Computer Science(), vol 12155. Springer, Cham. https://doi.org/10.1007/978-3-030-52794-5_11
- [33] Leslie G. Valiant. "Universal Circuits (Preliminary Report)". In: Proceedings of the Eighth Annual ACM Symposium on Theory of Computing. STOC '76. Hershey, Pennsylvania, USA: Association for Computing Machinery, 1976, pp. 196–203. doi: 10.1145/800113.803649.
- [34] Jitendra Bhandari et al. Not All Fabrics Are Created Equal: Exploring eFPGA Parameters For IP Redaction. 2021. doi: 10.48550/ARXIV.2111.04222.
- [35] Gaurav Kolhe et al. "Breaking the Design and Security Trade-off of Look-up-Table-Based Obfuscation". In: ACM Trans. Des. Autom. Electron. Syst. 27.6 (June 2022). doi: 10.1145/3510421.
- [36] Gaurav Kolhe et al. "Security and Complexity Analysis of LUT-based Obfuscation: From Blueprint to Reality". In: 2019 IEEE/ACM ICCAD. 2019, pp. 1–8. doi: 10.1109/ICCAD45719.2019.8942100.
- [37] K. Xiao and M. Tehranipoor. "BISA: Built-in self-authentication for preventing hardware Trojan insertion". In: 2013 IEEE International Symposium on Hardware-Oriented Security and Trust (HOST). 2013, pp. 45–50. doi: 10.1109/HST.2013.6581564.
- [38] D. Sisejkovic and R. Leupers, "Trustworthy Hardware Design with Logic Locking," 2021 IFIP/IEEE 29th VLSI-SoC, Singapore, Singapore, 2021, pp. 1-2, doi: 10.1109/VLSI-SoC53125.2021.9606998.
- [39] K. Xiao, D. Forte, and M. Tehranipoor. "A Novel Built-In Self-Authentication Technique to Prevent Inserting Hardware Trojans". In: IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 33.12 (2014), pp. 1778–1791. doi: 10.1109/TCAD.2014.2356453.
- [40] Richard Jarvis and Michael McIntyre. Split manufacturing method for advanced semiconductor circuits. U.S. Patent no 20040102019A1, May 2004.
- [41] Yajun Yang et al. "How Secure Is Split Manufacturing in Preventing Hardware Trojan?" In: 25.2 (Mar. 2020). doi: 10.1145/3378163
- [42] Jeyavijayan Rajendran et al. "Security Analysis of Integrated Circuit Camouflaging". In: Proceedings of the 2013 ACM SIGSAC CCS '13. Berlin, Germany: Association for Computing Machinery, 2013, pp. 709–720. doi: 10.1145/2508859.2516656.
- [43] Satwik Patnaik et al. "Obfuscating the Interconnects: Low-Cost and Resilient FullChip Layout Camouflaging". In: Proceedings of the 36th ICCAD '17. Irvine, California: IEEE Press, 2017, pp. 41–48.
- [44] Dominik Sisejkovic. "Designing trustworthy hardware with logic locking". Veröffentlicht auf dem Publikationsserver der RWTH Aachen University; Dissertation, Rheinisch-Westfälische Technische Hochschule Aachen, 2022. Online-Ressource : Illustrationen. doi: 10.18154/RWTH-2022-02625.
- [45] M. R. Fadiheh *et al.*, "An Exhaustive Approach to Detecting Transient Execution Side Channels in RTL Designs of Processors," in *IEEE Transactions on Computers*, vol. 72, no. 1, pp. 222-235, 1 Jan. 2023, doi: 10.1109/TC.2022.3152666.
- [46] IEEE Spectrum, "Stopping Hardware Trojans in Their Attacks", <https://spectrum.ieee.org/stopping-hardware-trojans-in-their-tracks> Accessed: Nov, 2023.

Lennart M. Reimann is the chief engineer at the Institute for Communication Technologies and Embedded Systems at RWTH Aachen University.

Dominik Sisejkovic is a research activity manager for cybersecurity automation at Corporate Research, Robert Bosch GmbH, Hildesheim, Germany.

Rainer Leupers is a professor at the Institute for Communication Technologies and Embedded Systems at RWTH Aachen University.



While computing systems rely on hardware as a "root-of-trust" for security, new vulnerabilities exist that cannot be patched in software alone. Security verification and the development of defense mechanisms must be pursued at the microarchitectural hardware level, utilizing formal methods specifically tailored for security verification.

Microarchitectures as Root-of-Trust in Computing Systems – Research Needs in Formal Security Analysis

by Wolfgang Kunz and Dominik Stoffel

System-on-Chips (SoCs) and embedded systems are ubiquitous in modern society. With their abundance of connectivity features they create a new attack surface for cyberattacks. Our trust in computing systems depends mainly on the provided safety and security features of the underlying computing hardware. Although the majority of the advanced security features, such as end-to-end encryption, are implemented at the software level, they rely on basic hardware primitives to deliver the intended functionalities. In common terminology, such hardware primitives form the “*root-of-trust*” of the computing system. They constitute a set of trusted functionalities to ensure the security of the system. Any security flaw in the hardware root-of-trust can affect virtually all applications deployed on the system.

Hardware systems are difficult or, in some cases, even impossible to patch, which exacerbates the challenge of dealing with hardware security flaws. Countless reports in recent years on system vulnerabilities at the hardware level, e.g., [1], [2], attest to the fact that hardware security flaws can pose a genuine threat to the overall system security. The Common Weakness Enumeration database (CWE) [3] has acknowledged this problem by

including hardware vulnerabilities as a separate category of security weaknesses.

The role of hardware in system security is not limited to providing security-related features to support software functions. Weaknesses in the hardware design itself can introduce severe vulnerabilities to the computing system. At the microarchitectural level, these weaknesses mostly have two sources: the hardware circuit executing a security-critical software application may leak confidential information through side channels, in particular timing of the software execution [4], or (possibly very subtle) design bugs escape conventional verification procedures and cause security risks for the entire system.

“State of computer security is embarrassing for all of us in the computing field. It seems unlikely systems will ever become secure using software-only solutions”.

David Patterson,
Keynote IEEE/ACM Design Automation Conference, 2018

This article explores using formal methods for verifying hardware security, a shift from the traditional software-centric approach. Historically, cybersecurity focused on software vulnerabilities,

but the 2018 discovery of hardware-based attacks like Spectre and Meltdown, exploiting ISA-invisible side channels such as transient execution side channels (TES), revealed the limits of software-only defenses. Subsequent attacks (e.g., [3], [4]) have further challenged the effectiveness of existing solutions and underscore the need for more comprehensive hardware security strategies.

“The potential for a transient execution to extract data being carried across a branch or a load is still a new field of research. Even though transient execution attacks are highly complex and difficult to carry out successfully outside of a lab, we expect it to remain a persistent focus area for researchers and the computer industry.”

Martin Dixon, VP for Security at Intel, 2021

Today, re-establishing trust in the microarchitectures of computing systems has become one of the main goals in the computer industry and among chip makers. Microarchitectural descriptions at the Register Transfer Level (RTL) are the point of reference for sign-off verification before the tape-out of a chip for manufacturing. They typically serve as the *golden model* of an SoC and are the basis for all design refinements at lower levels as well as for manufacturing.

Security verification and the development of defense mechanisms at the microarchitectural level have become rapidly growing research fields. There is general conviction that the formidable patch-and-pray cycles can only be overcome if comprehensive security guarantees are already provided during design and before tape-out. Formal verification bears promise to provide such guarantees. However, currently available techniques are tailored towards general functional design aspects and suffer from severe limitations when targeting microarchitectural security and side channels.

Key insights

- The immense attack surface exposed by microarchitectural security flaws and the vastness and diversity of SoC deployment domains necessitates extending formal methods from just functional correctness to systematic non-functional security verification. Research is vital to understand and formalize these diverse security risks and threats.

- Conventional functional verification applied to individual design components cannot provide comprehensive security guarantees for extensive hardware systems interacting with software; vulnerabilities often arise from the integration of multiple components and specific hardware-software interactions.
- The current solutions for mitigating security vulnerabilities often involve ad hoc processes, which require significant design alterations and necessitate communication with software developers. Moreover, while advanced security measures have been suggested to address timing side channels, they substantially escalate the manual design workload and introduce considerable hardware overhead, highlighting a pressing need for more refined, systematic, and comprehensive strategies.

Key recommendations

- Invest in research and development of tools for exhaustive security analysis of SoC microarchitectures. Invest in the formalization of microarchitectural threat models as targets for a new generation of non-functional formal verification methods.
- Devise, implement and adopt new design and verification methodologies based on these new tools, pursuing verification-driven, secure-by-construction design providing system-wide threat coverage.
- Contribute to open-source and public-domain initiatives like the RISC-V ecosystem to demonstrate use of the new tools and methodologies and to facilitate their widespread adoption and dissemination.

Security Vulnerabilities

Security analysis must address a wide spectrum of potential hardware vulnerabilities. At the microarchitectural hardware level, there are two main categories: *security-violating design bugs* and *microarchitectural timing side channels*.

Security-violating design bugs are the subset of all design bugs that, besides violating the functional specification, violate a relevant security target. In principle, bugs can be detected by conventional functional verification. However, this requires the complete and correct specification of the entire design as well as of all its security mechanisms.

Substantial effort is demanded from verification engineers to cover all security-relevant functional behaviors by a set of properties specified in a property specification language such as SVA (SystemVerilog Assertions). Even then, there may still be verification gaps remaining: security issues related to the communication between modules or to the interaction between hardware and firmware are easily missed [5].

A design that is bug-free may still be vulnerable to *side channels*. At the microarchitectural level, side channels are based on *timing*. Although a program may not have access rights to a certain set of data, depending on this data, one and the same program may behave slightly differently in terms of its own computation results, i.e., what data it stores in which registers and at which time points. These differences only affect the detailed timing of the microarchitectural implementation and have no impact at the level of the *instruction set architecture* (ISA), i.e., they do not affect the correct functioning of the program as observed by the programmer. However, if these subtle alterations of the program's execution at the microarchitectural level are caused by secret data to which the program must not have access, this may open a "side channel". An attacker, owning (and creating) such a program, may trigger and observe these alterations to infer secret information. This is called a "microarchitectural side channel attack".

The following observation is key to classifying microarchitectural side channels and the corresponding verification targets (properties): In the conventional attack scenario the attacker process by itself is not capable of controlling both ends of a side channel. To steal secret information, it must interact with another process initiated by the system, the "victim process", which manipulates the secret and "makes a noise". Common defenses at the software level include constant-time encryption [6] and cache access pattern obfuscation [7]. They prohibit the information flow at the "sending end" of the channel, i.e., the one owned by the victim process.

Although securing encryption software against these attacks is challenging because it demands a deep understanding of microarchitectural details, in the past, the threat of microarchitectural side channels was generally perceived to be limited to a small set of software applications. This general intuition, however, was drastically changed by the discovery of *transient execution side channel (TES) attacks*. Even though they use similar channels for

exfiltrating information, TES attacks are fundamentally different from classical microarchitectural side channels. TES attacks exploit side effects of transient instruction execution, a phenomenon not visible in the sequential execution semantics of the ISA. Transient execution occurs when the processor speculatively executes instructions ahead of time that it needs to roll back if the speculation turns out to be wrong. Such transient instruction execution may leak secret data through timing side channels and is the root cause for TES attacks. Without affecting the ISA-level results of the program, the attacker triggers transient executions of instructions that depend on secret data. In this way, the attacker does not rely on a vulnerability within a victim software to make a noise. In fact, the TES attacker controls both ends of the channel, the part that triggers the side effect and sends out the information as well as the part that observes it. This makes TES attacks more threatening than the earlier known timing side channels. In the TES scenario, a single user-level attacker program can establish a microarchitectural side channel leaking parts of the memory which are not accessed by any other program. Such HW covert channels not only can destroy the usefulness of encryption and secure authentication schemes but can steal data from essentially anywhere in the system. As a result, unlike classical side channel attacks, TES attacks threaten the overall security of the system and its root of trust.

Since the publication of the Spectre [8] and Melt-down [9] attacks in 2018, numerous new TES-based attacks have been discovered (e.g., MDS attacks [10], [11], [12]), speculative store bypass [13], speculative interference [14]), with many of them targeting a previously patched system (e.g., Fallout attack [10]), calling for new attention towards hardware security.

Needed: A Formalization of Threat Models

When considering both, the vast scope of the possible microarchitectural security threats, and the great variety of deployment domains for SoCs, it becomes apparent that the attack surface is huge. Such a challenge can only be met if formal methods are extended beyond traditional functional correctness checking to non-functional security verification. This must be done in a systematic way such that the relevant threats in their large diversity can be described, and many different use scenarios for computing systems can be covered.

Research is needed to characterize security risks and to develop and formalize the threats. In other words, we need a *taxonomy of threat models*. A threat model for hardware captures the *security targets* for a system in combination with a *profile of the attackers*.

Security targets for hardware most commonly are requirements of *confidentiality* or *integrity*. *Confidentiality* of hardware is given if all information stored or processed in the system is protected against being retrieved by an unauthorized entity. *Integrity* means preventing an attacker from changing or influencing a part of the system that is specified as protected.

The attacker profile makes assumptions about how attackers can access the system and what methods they can use to exploit potential vulnerabilities of the system. For example, an attacker may access the system by running an unprivileged user process. Another threat model may consider access to the security-critical system through a third-party IP which is added to the system and which the attacker controls. Specific threat models are the basis for the specification of verifiable properties. The challenge consists in formulating these properties in such a way that a large spectrum of different threat models is covered by a manageable set of properties.

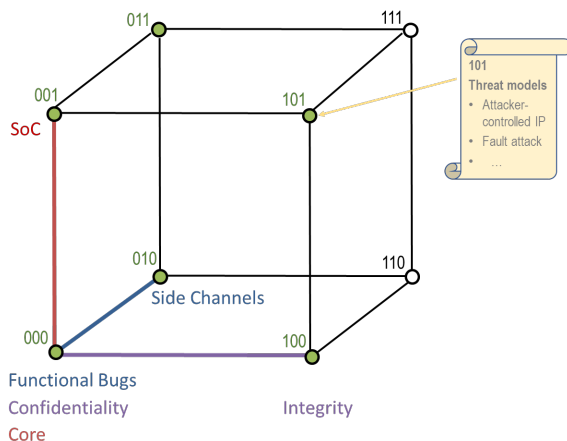


Figure 1: Space of threat models

Figure 1 illustrates the space of threat models that must be analyzed for specifying verifiable properties. The points associated with important threat models are marked in green color. Firstly, it is meaningful to distinguish between security-violating functional bugs and non-functional

vulnerabilities such as side channels. This is the blue dimension in the shown space.

Secondly, the vertical dimension (red color) of the cube in

Figure 1 distinguishes between vulnerabilities that occur only in *cores* and those requiring a global analysis of the entire *System-on-Chip (SoC)*. For example, TESs, such as Spectre and Meltdown, only require an analysis of the core while the root cause of other types of timing side channels can be distributed over several locations of the SoC.

Thirdly, the horizontal axis of the cube separates threat models related to the security target of *confidentiality* from those related to *integrity*.

We give some examples of threat models related to specific points of the cube in

Figure 1:

Threat Model 010

- *Security target:* Confidentiality of data in protected memory locations
- *Attacker Profile:* Attacker can run any program on the core with user-level privileges.
- *Class of vulnerabilities:* Transient Execution Side Channel in cores

Note that several threat models can belong to each point in the space of

Figure 1, for example:

Threat Model 101a

- *Security target:* Integrity of information and integrity of operation in security-critical parts of SoC
- *Attacker Profile:* Attacker controls a third-party IP which communicates with the security-critical SoC domain.
- *Class of vulnerabilities:* Design bugs (insufficient protection mechanisms)

Threat Model 101b

- *Security target:* Integrity of information and integrity of operation in security-critical parts of SoC

- *Attacker Profile*: Attacker can inject faults by laser light anywhere in the SoC.
- *Class of vulnerabilities*: Design bugs (insufficient protection mechanisms)

Hardware security engineers conduct threat analyses to create relevant threat models for specific designs and deployment domains. These models form the basis for verification engineers to define verification targets as formal properties, specified in languages like SVA. It's crucial to globally formalize threat models, aiming to cover a broad spectrum of vulnerabilities and attacker profiles, including "unknown unknowns", i.e., yet undiscovered vulnerabilities. The overall goal is to create generic, re-usable verification IPs.

Needed: New Tools

Late detection of hardware security flaws can incur tremendous costs, calling for *new functional verification* techniques specifically for hardware security. Rather than being exhaustive with respect to a complete functional design specification, the new tools must be *exhaustive with respect to a well-defined threat model* and deliver well-defined security guarantees. The new tools must be scalable and, at the same time, amenable to adoption by current industrial hardware design flows.

While conventional functional verification of a design, in principle, also avoids security-critical bug escapes, the manual effort for such an exercise is prohibitively large. The state-of-the-art flow for security verification based on functional property checking is shown in the left part of Figure 2. The targeted security features (Box II) typically result from a high-level, architectural perspective. The design specification is extended by an additional *functional* specification (Box III) of these security features which guides their integration into the RTL implementation. This is followed by functional verification procedures (Box IV). It turns out, however, that this classic approach is not always sufficient. Not only does conventional functional verification miss side channels, but also the abstract security requirements can be extremely difficult to map to functional specifications, requiring a detailed, microarchitectural understanding of security threats. Therefore, the specification itself (Box II) may fail to cover relevant aspects of the global threat model. Experience shows that, even when choosing security features conservatively, the conventional design process can miss subtle, yet hazardous security gaps and gives rise to the widely spread complaint about a never-ending "patch-and-pray" cycle.

Therefore, new approaches to formal hardware security verification are needed that target security

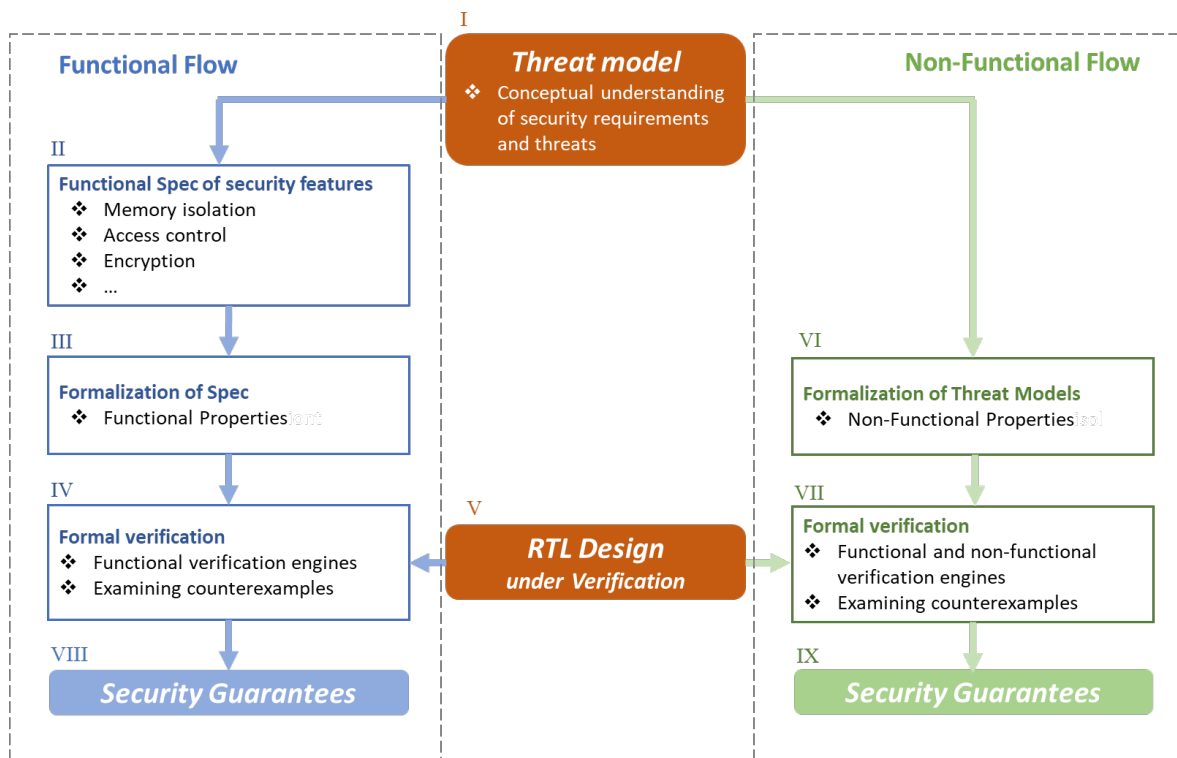


Figure 2: Functional vs. Non-functional verification flow for security

properties directly. Such a flow is shown on the right side of Figure 2. Instead of developing a detailed (and error-prone) functional specification, this approach starts from the applicable threat model (Box I) and formalizes security requirements rather than detailed functional behaviors. This leads to specifying non-functional security properties (Box VI) which are orthogonal to conventional functional specifications. Since these properties directly target global security requirements without the need of a functional specification for the intended defense mechanisms, these methods have the potential to cover security breaches (both functional breaches and side channels) which are easily missed by the conventional approach. Both conventional solvers for functional verification as well as specialized solvers, such as [15], can serve as a basis to extend formal hardware verification for such non-functional targets (Box VII).

In addition to functional or explicit information leakages that violate the security requirements, implicit information flows through *timing side channels* must also be considered. This complicates the verification process significantly because the functional specification, which is untimed, cannot make such a requirement. Therefore, there is a lack of proper specification techniques for security against timing side channels.

New formal tools are needed which target the threats by timing side channels in a systematic way. This research can build upon initial successes in academia to detect transient execution side channels but must extend the scope to all other side channels relevant at the microarchitectural level. The new methods must be capable of handling different architectures ranging from simple in-order processors without speculation to out-of-order processors with speculation.

Needed: New Flows and Methodologies

Running a single tool on a specific design or design component can never lead to global security guarantees formulated for a large hardware system and its interface with software. Instead, such guarantees must result from a security-driven flow that combines the results of different methods and tools across components and design layers. The needed extensions to today's flows have a "horizontal" and a "vertical" dimension.

The horizontal dimension concerns the structural composition of hardware designs consisting of multiple interconnected modules. Many security issues are introduced into the system through the *integration* of these components, and a vulnerability in one component may only be exploited through its communication with other components. Detecting such vulnerabilities requires analyzing information flows across multiple components which is usually a computationally expensive task for formal verification techniques.

This calls for new scalable verification *methodologies* exploiting specific advantages of different methods to cover system-wide security for a given threat model. The new methodologies are required to combine the results of different tools to compose global guarantees on the entire system. Formal verification in such a setting can be based on sound abstraction techniques over different stages of the design flow.

The vertical dimension concerns the Hardware/Software Interface. Many hardware security issues are only exposed if they are triggered by a specific interaction between hardware and software. Formal verification of such mechanisms demands proper modeling of hardware/software interaction.

Needed: Verification-Driven/Secure-by-Construction Design

Experience in industry and academia shows that most SoC hardware designs suffer from numerous security flaws based on both microarchitectural timing side channels and functional design bugs. Fixing design bugs is usually an ad hoc process which solves the problem by making design changes and/or communicating possible restrictions for the software layer with the software developers. Fixing timing side channels is a more demanding procedure. Advanced security features have been proposed, such as those based on information flow tracking [16], [17], that promise effective measures against these vulnerabilities. However, this comes with significant costs: the manual RTL design effort increases drastically, and the new architectures come with a significant hardware overhead that so far has only been estimated based on abstract system simulations. Only a few RTL architectures with security guarantees for side channels have been proposed.

Clearly, research is needed to explore new security architectures at the RTL. Formal security analysis can make a major contribution to developing new design methodologies leading to new security architectures. Formal verification precisely and exhaustively determines *all* attack scenarios that are relevant under the specified threat model. This knowledge can be very valuable to i) determine the root causes of the security weakness and ii) to derive fixes that avoid excessive conservatism. Note that, without the knowledge of formal tools, current measures for security often employ “blanket fixes” that cover a large (but not fully understood) spectrum of weaknesses.

References

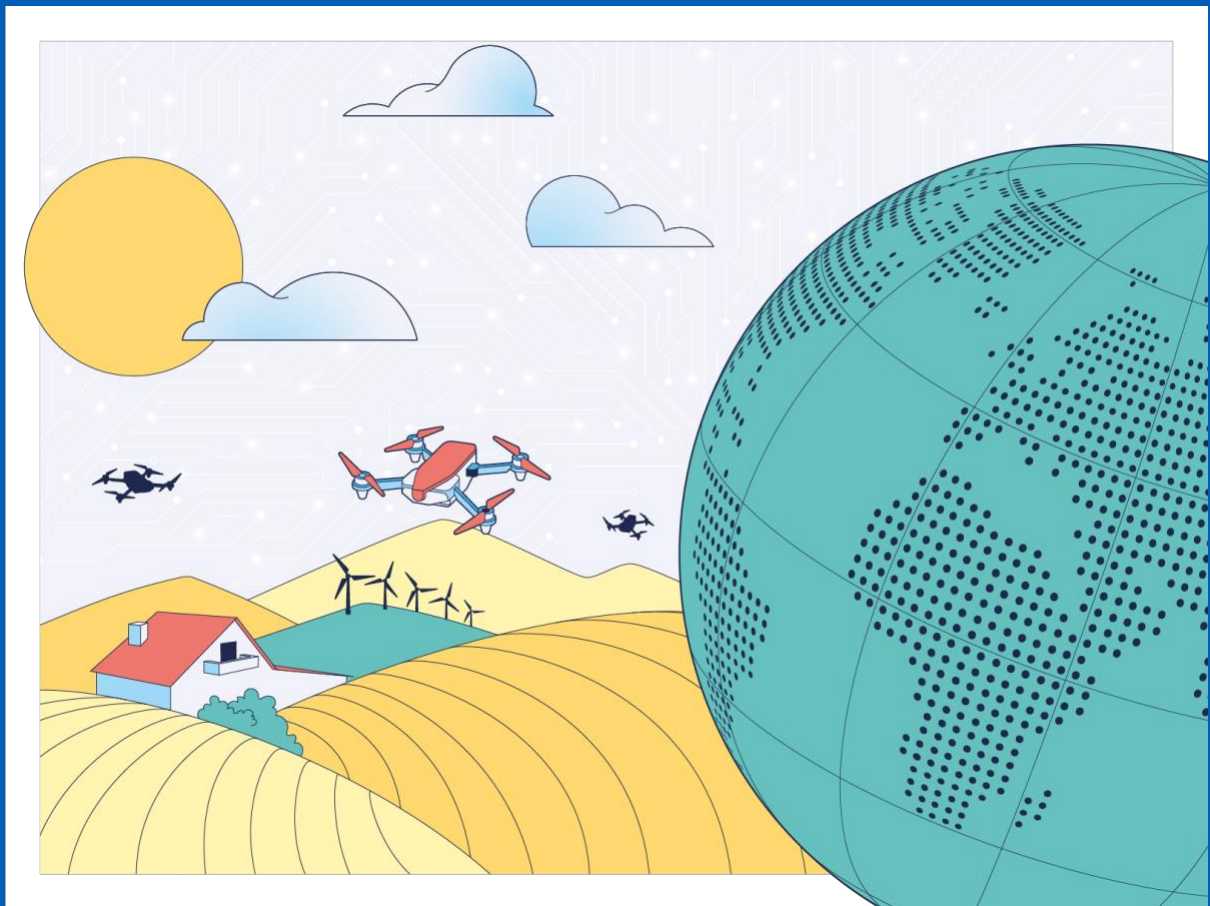
- [1] S. J. Vicarte, M. Flanders, R. Paccagnella, G. Garrett-Grossman, A. Morrison, C. Fletcher and D. Kohlbrenner, "Augury: Using data memory dependent prefetchers to leak data at rest," *IEEE Symposium on Security and Privacy (SP)*, p. pp. 1518–1518, 2022.
- [2] M. Gross, N. Jacob, A. Zankl and G. Sigl, "Breaking trustzone memory isolation through malicious hardware on a modern FPGA-SoC," in *Proceedings of the 3rd ACM Workshop on Attacks and Solutions in Hardware Security Workshop*, 2019.
- [3] "Common Weakness Enumeration," [Online]. Available: <https://cwe.mitre.org/>.
- [4] Y. Y. a. K. Falkner, "FLUSH+ RELOAD: A High Resolution, Low Noise, L3 Cache Side-Channel Attack," *USENIX Security Symposium*, p. 22–25, Vol. 1, 2014.
- [5] G. Dessouky, D. Gens, P. Haney, G. Persyn, A. Kanuparthi, H. Khattri, J. M. Fung, A.-R. Sadeghi and J. Rajendran, "Hardfais: Insights into software-exploitable hardware bugs," in *USENIX Security Symposium*, 2019.
- [6] D. Jayasinghe, R. Ragel and D. Elkaduwe, "Constant time encryption as a countermeasure against remote cache timing attacks," in *IEEE 6th International Conference on Information and Automation for Sustainability (ICIAIS)*, 2012.
- [7] J. Kong, O. Acicmez, J.-P. Seifert and H. Zhou, "Deconstructing new cache designs for thwarting software cache-based side channel attacks," in *Proceedings of the 2nd ACM workshop on Computer security architectures*, 2008.
- [8] P. Kocher, D. Genkin, D. Gruss, W. Haas, M. Hamburg, M. Lipp, S. Mangard, T. Prescher, M. Schwarz and Y. Yarom, "Spectre attacks: Exploiting speculative execution," *arXiv preprint arXiv:1801.01203*, 2018.
- [9] M. Lipp, M. Schwarz, D. Gruss, T. Prescher, W. Haas, S. Mangard, P. Kocher, D. Genkin, Y. Yarom and M. Hamburg, "Meltdown," *arXiv preprint arXiv:1801.01207*, 2018.
- [10] C. Canella, D. Genkin, L. Giner, D. Gruss, M. Lipp, M. Minkin, D. Moghimi, F. Piessens, M. Schwarz, B. Sunar, J. Von Bulck and Y. Yarom, "Fallout: Leaking Data on Meltdown-resistant CPUs," in *Proc. ACM Conference on Computer and Communications Security (CCS)*, 2019.
- [11] M. Schwarz, M. Lipp, D. Moghimi, J. Van Bulck, J. Stecklina, T. Prescher and D. Gruss, "ZombieLoad: Cross-privilege-boundary data sampling," *arXiv preprint arXiv:1905.05726*, 2019.
- [12] S. van Schaik, A. Milburn, S. Österlund, P. Frigo, G. Maisuradze, K. Razavi, H. Bos and C. Giuffrida, "RIDL: Rogue In-Flight Data Load," in *IEEE Symposium on Security and Privacy (S&P)*, 2019.
- [13] V. Kiriansky and C. Waldspurger, "Speculative buffer overflows: Attacks and defenses," *arXiv preprint arXiv:1807.03757*, 2018.
- [14] M. Behnia, P. Sahu, R. Paccagnella, J. Yu, Z. Zhao, X. Zou, T. Unterluggauer, J. Torrellas, C. Rozas, A. Morrison and others, "Speculative interference attacks: Breaking invisible speculation schemes," *arXiv preprint arXiv:2007.11818*, 2020.
- [15] K. v. Gleissenthall, R. G. Kıcı, D. Stefan and R. Jhala, "IODINE: Verifying Constant-Time Execution of Hardware," in *28th USENIX Security Symposium (USENIX Security 19)*, 2019.
- [16] J. Yu, M. Yan, A. Khyzha, A. Morrison, J. Torrellas and C. W. Fletcher, "Speculative taint tracking (STT) a comprehensive protection for speculatively accessed data," in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*, 2019.
- [17] K. Loughlin, I. Neal, J. Ma, E. Tsai, O. Weisse, S. Narayanasamy and B. Kasikci, "DOLMA: Securing Speculation with the Principle of Transient Non-Observability," in *30th USENIX Security Symposium (USENIX Security 21)*, 2021.
- [18] Y. Yarom, D. Genkin and N. Heninger, "CacheBleed: a timing attack on OpenSSL constant-time RSA," *Journal of Cryptographic Engineering*, vol. 7, p. 99–112, 2017.
- [19] C. Percival, "Cache missing for fun and profit," in *BSDCan*, 2005.
- [20] D. Gullasch, E. Bangerter and S. Krenn, "Cache games—Bringing access-based cache attacks on AES to practice," in *IEEE Symposium on Security and Privacy (SP)*, 2011.
- [21] A. Purnal, F. Turan and I. Verbauwhe, "Prime+ Scope: Overcoming the Observer Effect for High-Precision Cache Contention Attacks," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021.
- [22] P. Pessl, D. Gruss, C. Maurice, M. Schwarz and S. Mangard, "DRAMA: Exploiting DRAM Addressing for Cross-CPU Attacks," in *25th USENIX Security Symposium (USENIX Security 16)*, 2016.
- [23] O. Acicmez and J.-P. Seifert, "Cheap hardware parallelism implies cheap security," in *Workshop on Fault Diagnosis and Tolerance in Cryptography (FDTC)*, 2007.
- [24] P. C. Kocher, J. Jaffe and B. Jun, "Differential Power Analysis," *Advances in Cryptology*, pp. 388–397, 1999.
- [25] Y. Liu, L. Wei, Z. Zhou, K. Zhang, W. Xu and Q. Xu, "On code execution tracking via power side-channel," *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, p. 1019–1031, 2016.
- [26] O. Reparaz, B. Bilgin, S. Nikova, B. Gierlichs and I. Verbauwhe, "Consolidating Masking," *CRYPTO*, 2015.
- [27] D. Jayasinghe, A. Ignjatovic, J. A. Ambrose, R. Ragel and S. Parameswaran, "QuadSeal: Quadruple algorithmic symmetrizing countermeasure against power based side-channel attacks," *International Conference on Compilers, Architecture and Synthesis for Embedded Systems (CASES)*, pp. 21–30, 2015.

- [28] V. Arribas, S. Nikova and V. Rijmen, "VerMI: Verification Tool for Masked Implementations," *IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, pp. 381-384, 2018.
- [29] R. Bloem, H. Gross, R. Iusupov, B. Könighofer, S. Mangard and J. Winter, "From Verification of Masked Hardware Implementations in the Presence of Glitches," *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 321-353, 2018.
- [30] ARM TrustZone Technology.
- [31] D. Lee, D. Kohlbrenner, S. Shinde, K. Asanović and D. Song, "Keystone: An Open Framework for Architecting Trusted Execution Environments," in *EUROSYS*, 2020.
- [32] "The seL4@ Microkernel," [Online]. Available: <https://sel4.systems/>.
- [33] A. Ferraiuolo, M. Zhao, A. C. Myers and G. E. Suh, "HyperFlow: A processor architecture for nonmalleable, timing-safe information flow security," in *ACM SIGSAC Conf. on Computer & Communications Security*, 2018.
- [34] RISC-V Foundation, The RISC-V Instruction Set Manual, Volume II: Privileged Architecture, Version 1.10, A. Waterman and K. Asanović, Eds., 2017.
- [35] R. Baranowski, M. A. Kochte and H.-J. Wunderlich, "Reconfigurable Scan Networks: Modeling, Verification, and Optimal Pattern Generation," *ACM Trans. Des. Autom. Electron. Syst.*, vol. 20, pp. 30:1-30:27, March 2015.
- [36] G. Cabodi, P. Camurati, F. Finocchiaro and D. Vendramineto, "Model Checking Speculation-Dependent Security Properties: Abstracting and Reducing Processor Models for Sound and Complete Verification," in *Intl. Conf. on Codes, Cryptology, & Information Security*, 2019.
- [37] G. Cabodi, P. Camurati, S. F. Finocchiaro, F. Savarese and D. Vendramineto, "Embedded Systems Secure Path Verification at the HW/SW Interface," *IEEE Design & Test*, vol. 34, p. 38-46, 2017.
- [38] P. Subramanyan and D. Arora, "Formal verification of taint-propagation security properties in a commercial SoC design," in *Design and Test In Europe (DATE)*, 2014.
- [39] M. R. Clarkson and F. B. Schneider, "Hyperproperties," *Journal of Computer Security*, vol. 18, p. 1157-1210, 2010.
- [40] M. R. Fadiheh, D. Stoffel, C. Barrett, S. Mitra and W. Kunz, "Processor Hardware Security Vulnerabilities and their Detection by Unique Program Execution Checking," in *IEEE Design Automation and Test in Europe (DATE)*, 2019.
- [41] M. R. Fadiheh, A. Wezel, J. Mueller, J. Bormann, S. Ray, J. M. Fung, S. Mitra, D. Stoffel and W. Kunz, "An Exhaustive Approach to Detecting Transient Execution Side Channels in RTL Designs of Processors," in *preview of IEEE Transactions on Computers*, 2022.
- [42] J. Müller, M. R. Fadiheh, A. L. Duque Anton, T. Eisenbarth, D. Stoffel and W. Kunz, "A Formal Approach to Confidentiality Verification in SoCs at the Register Transfer Level," in *IEEE/ACM Design Automation Conference (DAC)*, 2021.
- [43] M. Goli and R. Drechsler, "Early SoCs Information Flow Policies Validation Using SystemC-Based Virtual Prototypes at the ESL," *ACM Transactions on Embedded Computing Systems*, <https://doi.org/10.1145/3544780>, 2022.
- [44] R. Guanciale, M. Balliu and M. Dam, "InSpectre: Breaking and fixing microarchitectural vulnerabilities by formal analysis," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 2020.
- [45] S. A. Seshia and P. Subramanyan, "UCLID5: Integrating modeling, verification, synthesis and learning," in *2018 16th ACM/IEEE International Conference on Formal Methods and Models for System Design (MEMOCODE)*, 2018.
- [46] C. Trippel, D. Lustig and M. Martonosi, "CheckMate: Automated synthesis of hardware exploits and security litmus tests," in *2018 51st Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2018.

Wolfgang Kunz and Dominik Stoffel are professors in the Department of Electrical and Computer Engineering at University of Kaiserslautern, Germany.

Sponsored by the German Agentur für Innovation in der Cybersicherheit GmbH.

Sustainability





Sustainability is the biggest challenge of the 21st century; without it, human civilization will be unable to continue. This chapter explores the relationship of computing technologies to sustainability.

The race for sustainability

by Koen De Bosschere

Climate change is now felt all over the planet, leading to changing natural ecosystems, and eventually to loss of biodiversity. Climate change is not only about melting ice caps, and rising sea levels, but also about fresh drinking-water supplies, food security, and water needed for manufacturing, agriculture, transportation and energy production. Climate change affects many aspects of our daily lives.

However, sustainability is wider than climate change. It is about organizing the world within the limits of planet Earth both today and for the generations to come. One aspect of it is the use of raw materials, which are finite on a finite planet. Modern information and communication technology (ICT) devices require more than 60 minerals from all over the world, some of which are hard to mine, are scarce, or are mined in socially unacceptable conditions. Once used in a device, they are difficult to recycle; but without efficient reuse and recycling, there cannot be true sustainability.

Modern ICT devices are very power efficient, leading to a situation where more energy is needed to produce them than they will use during their entire lifetime (“embodied” versus “operational” energy use). Hence, integrating multiple functions into one device is more sustainable than creating separate devices per function. Contrary to common belief, continuing to use an old device is often more sustainable than buying a new, more power-efficient one. This shows that a full lifecycle assessment can lead to counterintuitive conclusions, and that such an assessment is essential in order to come up with sustainability solutions with real impact.

Sustainability is perhaps the most important grand challenge of the 21st century and it affects everything we do. This chapter looks into ways to make computing sustainable and suggests how computing can contribute to sustainability goals.

This part contains three contributions.

- **“What does it mean to be sustainable?”** This article discusses the definition of sustainability in the context of ICT: energy, water and material consumption, and how to reduce and/or neutralize them. The main conclusion is that there is a need for universally agreed life cycle models for computing devices and services. Without such models, it is impossible to make effective sustainable design decisions, to make operational decisions when they are in use, and design new business models for the ICT industry.
- **“Sustainable materials and production”.** This article focusses on the environmental impact of the physical layer of computing systems. It analyzes the origin of the embodied emissions of modern technology nodes: the high number of required semiconductor grade minerals and the increasing number of energy, gas, and water consuming process steps to produce a chip. Finally, there is the technical challenge to effectively recycle semiconductor components.
- **“Towards sustainable computer architecture: A holistic approach”.** This article argues that sustainable development requires a holistic approach and involves multi-perspec-

tive thinking. Applied to computing, sustainable development means that we need to consider the entire lifecycle of a product. Analyzing current trends reveals that the embodied energy footprint is, or will soon be, more significant compared to the operational energy footprint. The article summarizes what computer architects and engineers can and should do to better understand the environmental impact of computing, and to design sustainable computer systems.

for green applications”, and works collectively with various domains in investigating how ICT technology could reduce the environmental impact of key processes in these domains.

Key recommendations

- **Design and improve validated lifecycle models.** These models should not only encompass factors like embodied and operational energy but should also comprehensively account for the ecological impact of activities such as mining, water usage, the use of chemicals in production, and the environmental consequences at the end of a product’s life cycle.
- **Develop sustainability-focused design methodologies.** By making a low ecological footprint a core design objective, designers will naturally take into account factors like repairability, reusability potential, recyclability, and the effective management of end-of-life processing right from the initial stages of product development.
- **Develop sustainable production techniques for semiconductor components.** Such techniques should reduce the emission of greenhouse gasses, the use of toxic, hazardous and critical raw materials, the consumption of ultra-pure water, and the amount of energy needed to produce semiconductors.
- **Create new business models incorporating the full lifecycle.** HiPEAC underscores that the absence of viable sustainable business models is a significant barrier to meaningful actions by companies aimed at enhancing the sustainability of their products. Therefore, rethinking business models is crucial to drive positive change in the sustainability landscape.
- **Prioritize ICT for green applications.** In addition to developing more sustainable digital technologies, HiPEAC recommends that the community focuses on developing “ICT

Koen De Bosschere is a professor in the electronics department of Ghent University, Ghent, Belgium.



Sustainability is the single most important societal challenge of the 21st century, and computing turns out to be both a part of the problem and a part of the solution.

What does it mean to be sustainable?

by Koen De Bosschere and Patrick Blouet

When IT devices are on display in shops, they look shiny, attractive, clean. At the time of buying, very few people think about the impact of mining raw materials, the purification of the materials before they can be used in a fab, the energy the device will consume during its entire lifetime, and where it will eventually end up after they stop using it. Most users of the devices might not even be aware where and how the device they are about to buy has been produced. In order to become more sustainable, consumers have to be more knowledgeable about the environmental impact of computing devices, and this knowledge should be based on solid and universally accepted facts instead of beliefs. As long as some people disagree on the scientific facts and evidence, they might not take the right decision to reduce their carbon footprint, and they will have difficulties to assess the sustainability claims made by device producers.

Key insights

- Determining the sustainability of computing solutions turns out to be extremely difficult, especially if one wants to take into account the full life cycle of products, and the secondary effects in the sectors where they are applied.
- There are several sustainability beliefs that may hold at a local scale, but are not valid at the global scale, like efficiency = sustainability (does not take into account the rebound effect) or computing makes other sectors more sustainable (the so-called enablement, for

which there is little evidence over the last 50 years).

- Sustainability is a complex problem. The effective solutions must be based on universally accepted sustainability models, and not on popular beliefs.

Key recommendations

- There is an urgent need for practical and universally accepted life cycle models for computing devices and services. The models should be used to steer decisions in the design and operation of computing systems, placing it on par with other design objectives such as cost, power consumption, and security.
- There is a need to educate the computing community on sustainability and to encourage it to start working on it. The only antidote against greenwashing are consumers that have a basic understanding on what true sustainability actually means.

Introduction

It was Jean-Baptiste Colbert, one of the main ministers of the French king Louis XIV, who, 350 years ago, saw the economic potential of the production of marine timber in the French forests. Without the ordinance of 1669 drawn up under his leadership, the majestic French-style oak groves like those of Tronçais or Bercé in France, managed by several generations of foresters, would not be available to us today.

According to (Grober, 2007), the term “sustainability” was introduced in the 18th century and it referred to managing forests to ensure that future generations had access to enough wood for fuel and for timber. In the 20th century it was broadened beyond forests. The first report of the club of Rome, “The Limits to Growth” (D.H. Meadows et al., 1972), made very clear that exponential growth of resource consumption could not be sustainable on a finite planet, and that one day natural resources would be depleted. The same report already warned about the exponential increase of CO₂ in the atmosphere in 1972. Later, the definition of sustainability was broadened with social and economic aspects into the formulation of the “triple bottom line” consisting of people, planet and profit (Triple bottom line, 2008). Today’s generally accepted definition of sustainability are the 17 sustainable development goals defined by the United Nations in 2015 (The 17 Goals, 2015).

Of all 17 sustainability goals, *Goal 13: Climate Action* is probably the best known. It states that “To limit warming to 1.5°C, global greenhouse gas emissions must fall by 45% from 2010 levels by 2030 and to zero by 2050”. Already in 2018, the UN Secretary-General António Guterres warned the world that “Climate change is moving faster than we are”; since then, greenhouse gas emissions have only increased. Today, the *projected* greenhouse gas emissions will be 10% higher in 2030 than in 2010. In order to reach the goals of 2030, we have to reduce the emissions by 42% in the next seven years (United Nations Environment Programme, 2023). This will be next to impossible to realize. With the current efforts, the climate will warm up between 2.5-2.9°C by the end of the century. Such a temperature increase will have catastrophic consequences for the whole planet.

But why is it so difficult to decarbonize the world? Some countries do not care about climate change

because e.g. selling fossil fuels is a major source of income, or they simply do not have the resources to do it (e.g. in developing countries). Individuals are reluctant to cut down the use of fossil fuels because it requires investments (isolation, purchase of a heat pump or an electric car), or it requires lifestyle changes like moving to a smaller house, travel less, consume less, eat less meat, ... Some industries are reluctant to make the energy transition as it will impact their bottom line of the next quarter; unless forced to change, they prefer business-as-usual as long as possible. The consequence is that fossil fuel consumption is expected to further increase until around 2030, and that the gap between the required reduction and the actual consumption will increase until 2040 (SEI, Climate Analytics, E3G, IISD, and UNEP, 2023). In other words, that we are not moving in the right direction fast enough.

One of the reasons why it is very difficult to decarbonize the world, is that the world is literally built from carbon (Vaclav Smil, 2022). The built environment primarily depends on cement and steel, both of which require lots of energy to produce, energy which has been fossil fuel based for the last 200 years. Many of the objects that surround us are made from oil in the form of plastics: clothing, furniture, toys, packaging, bottles, ... The biomass that we grow (food, feed, wood, ...) needs fertilizer, most of which is made from natural gas. Hence, we do not only need fossil fuels because they generate hot air, but also because we need their molecules in the chemical industry.

Given the current trends, reducing the global greenhouse gas emissions by 42% by 2030 won’t be realistic anymore as illustrated in Figure 1. Even the global lockdown in the year 2020 created only a small dent in the emissions.

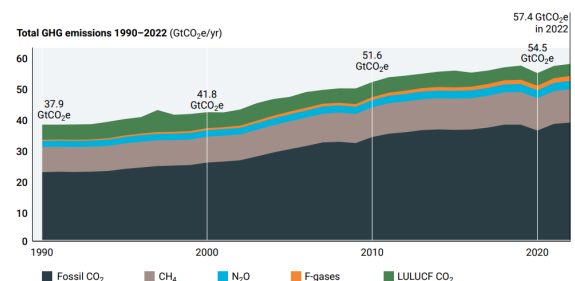


Figure 1. Total Green House gas emissions 1990-2022 (United Nations Environment Programme, 2023)

We simply cannot decarbonize a fossil fuel based global industrial ecosystem that took 200 years to

develop into a carbon-neutral one in few decades, especially because there are not yet proven alternatives for several types of emissions (e.g. air travel), and fossil fuel capital investments made today are expected to last for decades. For electricity production, there are several alternatives for fossil fuels (wind, solar, nuclear, hydro, biomass, ...), but given the investment rate, the targets for 2030 are practically unreachable.

In the meantime, climate change is accelerating (Figure 2), and the effects of the warming become noticeable in our daily lives in the form of more extreme weather events.

Experts warn that the use of fossil fuel not only emits CO₂, but also other particles like sulphur dioxide and soot that make the clouds more reflective and reflect up to 10% of the incoming radiation back into space. Phasing out the use of fossil fuels will automatically reduce the associated pollution (which is good), but also reduce global dimming, and thus increase the global temperature (which is not good).

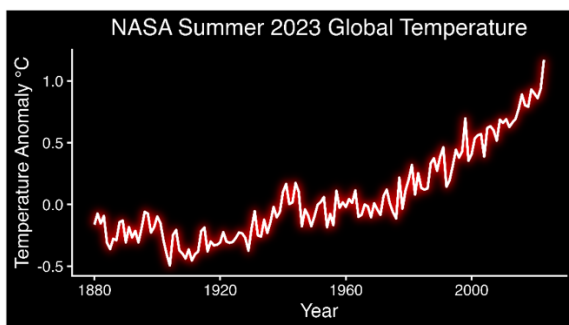


Figure 2. Temperature evolution 1880-2023 (NASA SVS | NASA Summer 2023 Temperature Media Resources)

So, even if we would stop burning fossil fuels tomorrow, the climate will continue to warm for some time due to the reduction of global dimming. No matter what we do to fight global warming, we will also have to adapt to the changing climate, and therefore, we need a two-pronged approach: reducing emissions as fast as we can, while preparing for climate change. The worst effects of climate change are the more violent weather phenomena (storms, droughts, floods), and their impact on food production in the northern hemisphere (where most of the global food production happens). E.g. a reduced runoff in the Colorado river due to less snowfall in the Rocky Mountains would have a serious impact on the agriculture in southern California, one of the food baskets of the USA. Similar scenarios apply to the whole

Mediterranean basin. A reduction of 10% in yield during a couple of years will lead to global famine, large scale migrations and political instability which does not help to increase food production (see Ukrainian war). Given the fact that agriculture is responsible for more than 25% of the global CO₂e emissions, it is paradoxical that in order to slow down global warming, the world should cut down on agriculture to avoid global famine in the future.

In any case, it is in our interest to find ways to adapt to climate change (e.g. by selecting crops that are better resistant to the new climatological situation and compatible with the local ecosystem, and to manage the rainfall in a more effective way by buffering it for future use or to let it infiltrate and replenish the water table, ...). It is mandatory for water to have several lives and to abandon the current model of single use.

In this article, we discuss some basic sustainability concepts.

Emissions by the IT industry: embodied vs. operational

Sustainability in the IT sector has two components. First of all, there are the resources needed to manufacture the electronic devices, and secondly, there are the resources needed to actively use the device, in practice it is primarily electrical energy and water. Although a single transistor or a single memory cell requires an almost infinitesimal amount of energy to switch, once multiplied by billions per device, and switching billions of times per second leads to measurable power consumption, ranging from less than 10 mW of operational energy for very low power devices to MWs of power for data centers. This multiplied by a few billion active devices on the planet leads to an industry consuming around 9.3% of the global electricity production and rising.

Concerning the minerals, modern IT devices are very complex devices that need complex supply chains to bring tens of semiconductor grade (= ultrapure) minerals from all continents together in one place, in order to integrate them into one device via a very sophisticated and energy and water consuming production process. The energy required and the emissions caused by this process are called the embodied energy and the embodied emissions, and they both contribute to global warming.

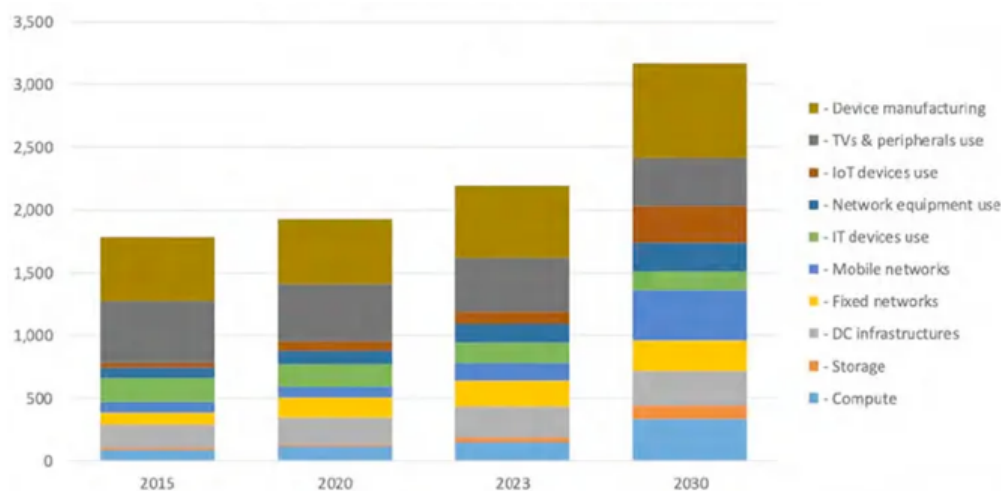


Figure 3. Evolution of IT energy demand

In order to contribute its share to the climate efforts, the IT industry should reduce its greenhouse gas emissions by 50% between now and 2030. However, according to Schneider Electric (Vincent Petit, 2021), the total energy consumption (used here as a proxy for the emissions) of the digital economy is projected to grow by 50% by 2030. So, instead of halving its energy footprint, it will double it!

A major source of growth is device manufacturing (the electricity part of the embodied energy), mobile networks (5G), and compute power (AI, blockchain, ...). The growth rate (2020-2030) in the operational energy will be more than 15% for storage, mobile networks and IoT devices. Where the share of energy consumption of the IT sector was 9.3% in 2023 (including the energy to manufacture devices) will grow to 11.5% in 2030. Unfortunately, the energy required for dismantling and end of life management is not taken into account in this study.

The carbon footprint of the IT industry has been growing by 1.8% per year during the last decades, and this is faster than the growth of the global emissions. Its carbon footprint in 2020 is 2.1-3.9% of the global greenhouse gas emissions, and 30% of it are embodied emissions.

Carbon neutral, net-zero, and carbon-negative

When reading sustainability announcements by companies it is important to carefully read the wording used in the announcement.

Carbon neutral means that an organization measures the CO₂ emissions caused by its operations, reducing them as far as possible, and offsetting the part that cannot be avoided (e.g. by reforestation, investing in renewable energy, or capturing CO₂ from the air). Any company or organization can fairly easily become carbon neutral by replacing fossil energy sources by renewable ones where possible and paying for the remaining CO₂ emissions. Even without sustainability investments, it is possible to become carbon neutral by offsetting more. For individuals, buying renewable electricity, driving an electric vehicle, installing a heat pump, and avoiding or offsetting air travel is a good start to become carbon neutral for the operational emissions. Making whole industries carbon neutral is however a totally different challenge.

The gold standard for organizations is however **net-zero**. This is more challenging as it requires that not only the carbon emissions are taken into account, but also other greenhouse gasses like methane, nitrous oxide, and fluorinated gas emissions. Some fluorinated gasses trap 1000 times more heat than CO₂. It is a common practice to convert all greenhouse gasses into CO₂ equivalents (CO₂e) to simplify the math.

Net-zero emissions do not only apply to the business operations themselves, but over the complete supply chain. This might be feasible in the IT industry, but much harder in e.g. agriculture, which is a major source of methane and nitrous oxide emissions.

Carbon-negative means that a particular device or process extracts more carbon from the air than it emits. Trees are carbon-negative which explains why they are used to offset carbon emissions.

Figure 3 shows that the IT sector will continue to grow, and to use more operational energy in the coming years (despite the fact that devices will continue to be more power efficient). This growth is faster than the rate at which the energy generation is decarbonized globally. This means that the IT related emissions will continue to increase until 2030.

In order to limit global warming to 1.5°C, we need carbon neutrality by 2050, and net-zero in 2070. This will require that all stakeholders work on their own targets, without exceptions. The intermediate goals in 2030 are an instrument to assess whether we are on track or not. According to (United Nations Environment Programme, 2023), we are not at all on track.

A case study: Apple watch and iPhone 15

Apple announced in 2023 that its latest Apple Watch is carbon neutral. Apple succeeded to reduce the lifetime carbon footprint of the watch by 75%, by using recycled materials, by avoiding air transportation, but especially by using renewable energy in manufacturing, and requiring its suppliers to do so too. It even decarbonized the electricity to power the watch during its full life cycle by investing in large-scale solar and wind projects. The remaining 25% that could not be avoided is offset in nature-based projects in South America. This announcement proves that it is possible to substantially reduce the lifecycle carbon footprint of a device, and to compensate for the rest, but it requires extra effort, and probably makes the device more expensive too.

On 13 September 2023, the iPhone 15 was launched. According to Apple, the total

lifetime emissions of the iPhone 15 Pro (128 GB) is 95 kg, but by using recycled and low-carbon materials and by using renewable energy in manufacturing it was able to reduce the carbon footprint by 26 kg CO₂e to 66 kg CO₂e. Of the 66 kg CO₂e, 83% goes into production, 3% in transportation, 15% in product use, and less than 1% in end-of-life processing. That means that the operational emissions are 15% of 66 kg CO₂e = 10 kg CO₂e, which means that more CO₂e (56 kg) was emitted during the production of the device (mining, fab, assembly, transportation, ...) than during its operation (10 kg). Even more surprising is that a device of only 171g causes 327 times more CO₂ emissions (56 kg) than its own weight.

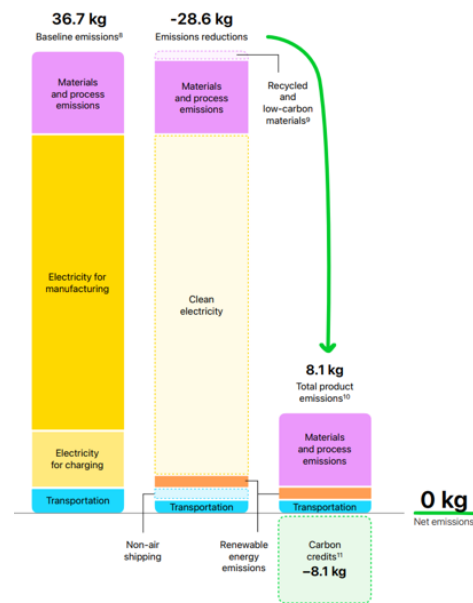


Figure 4. Carbon emissions Apple watch

Obviously, the operational emissions will depend on the usage intensity, and on the carbon intensity of the electricity used. The iPhone can play 23h of video on one battery charge of 12.7Wh. Assuming that the phone is recharged once per day, this will require 365 x 12.7 Wh = 4.6 kWh per year. With an average emission rate of 0.33 kg CO₂e per kWh, this adds up to 1.5 kg CO₂e per kWh which is very low compared to the 56 kg CO₂e emitted to produce the device. Since the total operational emissions are estimated to 10 kg CO₂e, Apple apparently assumes a lifetime of 10/1.5=6.6 years,

which is consistent with their recent announcement that they offer replacement parts for up to 7 years. That means that owning and using the iPhone creates an equivalent yearly emission of $66/6.6 = 10$ kg CO₂e per year. Extending the lifetime of the device from 6 to 7 years, leads a reduction of the yearly average emissions from 10.8 kg CO₂e to 9.5 kg CO₂e, that is a reduction of 1.3 kg CO₂e per year. That means that using it one year longer is about enough to offset the operational emissions for the device.

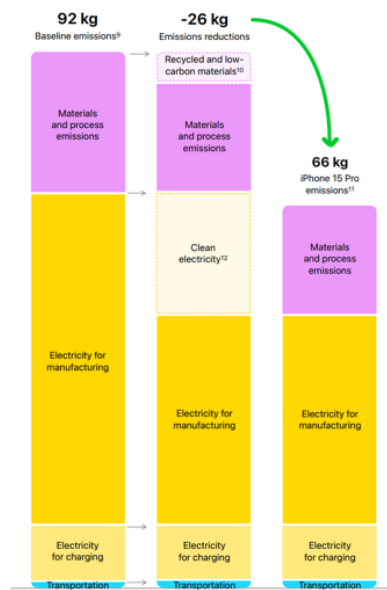


Figure 5. Carbon emissions iPhone 15

No matter how power-efficient the latest model of a smartphone is, the benefits of saving operational energy will never outweigh the energy required to build a new device. This is not only true for smartphones, but for all complex low-power devices like tablets and laptops. On top of this, the operational emissions will further decrease due to the increased use of renewable energy and due to the use of more power-efficient technology nodes. While reducing the operational energy consumption of a device is commendable, it's essential to evaluate whether the benefits in operational energy savings outweigh the additional embodied energy associated with advanced technology nodes (due to more advanced production equipment and processes). For instance, the greenhouse gas

emissions of a 3nm technology node wafer are double those of a 7nm node wafer. Over the years, the total emissions for the iPhone have stayed pretty stable: iPhone 13 (2021): 69 kgCO₂e, iPhone 14 (2022): 65 kg CO₂e, iPhone 15 (2023): 66 kg CO₂e. Similarly, when transitioning from a cloud-based solution to an edge-based solution, a comprehensive assessment of the total lifecycle emissions of both options is necessary to determine which one is more sustainable.

Clean electricity to the rescue?

As illustrated with the case of the Apple products, most of the emissions are electricity for (i) manufacturing the devices and (ii) for running the devices. On paper, such emissions can easily be avoided by using clean energy for the production and for the use of the devices. Unfortunately, reality is more complex. Unless the source of the energy is non-intermittent like hydro-electric energy, biomass or nuclear, there is the need to somehow store the energy to bridge the periods when the electricity production is low. Currently, the most reliable storage is the grid, which means that one sells the excess local renewable energy to the grid and buys the electricity back when the local production is too low (like during the night for solar). Obviously, this only works well if the grid company has storage capacity (batteries, pumped storage, ...) or can provide fossil-free electricity (nuclear, biomass, hydro-electric). If that is not the case, it will sell fossil-based electricity.

Hence, even if a facility like a data center comes with a renewable energy production system that produces (more than) the total yearly energy consumption, it does not mean that the consumption is 100% clean. It will most of the time be a mix of renewable and non-renewable energy sources. While this energy mix will be less carbon-intensive, it will not achieve carbon neutrality. The effect is even aggravated by the fact that the renewable production on the entire grid might at some point be higher than the consumption, leading to a situation in which the producer has to pay to inject energy into the grid, or that the grid company decides to shut down some installations to protect the grid, meaning that the renewable production system does not run at full capacity. Hence, unless the manufacturing and charging would be limited to the periods when there is enough clean

electricity, there will always be a carbon footprint that will be equal to the carbon footprint of the mix of energy sources that are used to power the grid at that time. Hence, replacing all electricity by clean electricity is challenging in practice.

Applying these clean energy insights rigorously may lead to inconvenient conclusions, such as the simplest way to reduce the carbon footprint of data centers being to relocate them to regions with the lowest average carbon intensity for electricity, which is likely in countries with abundant nuclear and hydroelectric power.

Carbon offsetting to the rescue?

The current practice to make products carbon neutral is to first reduce the emissions, and then to offset the emissions that cannot be avoided. A popular way to offset is investing in reforestation projects. A mature tree absorbs between 10-40 kg of CO₂ per year. In order to absorb 1000 kg of CO₂, a tree has to grow on average 40 years. Hence, offsetting by reforestation means that the CO₂ produced today (e.g. a flight) will be absorbed from the air during the next 40 years, i.e., by 2063. This means that the majority of the CO₂ offsetted today will keep contributing to global warming during the next 20 years (2043) until the trees are mature enough to fully absorb it. Hence, reforestation does not help to slow down global warming in the next decade, but it does in the future. This assumes that the trees do not die during that period (e.g. by storm, by wildfire, by drought, or by disease). Eventually the trees will die, or be cut, which means that at that moment the stored carbon will be released again. In order to avoid the release of the carbon, a tree that dies or is cut should be replaced by another tree to store the released carbon again. If not, carbon offsetting is only a way to temporarily store the carbon. Hence, carbon offsetting is at best a temporary solution until we find a scalable and affordable technical means to permanently store the captured carbon because the space for reforestation on the planet is finite and it has to compete with agriculture. The most important insight is that offsetting has no immediate effect on the release of CO₂ and cannot even guarantee that the CO₂ will be removed from the atmosphere in the future. Offsetting is basically a right to pollute and does not stop global warming in the short term. The only way to (immediately) neutralize emissions is to capture them and store them permanently (capture can happen at the source, or from the atmosphere).

Furthermore, companies that try to offset CO₂e emissions in order to reach net-zero are sometimes misleading themselves. Fluorinated gasses stay decades to centuries in the atmosphere and are removed by photodissociation in the stratosphere. There is no way to offset them by reforestation. Hence, converting them into CO₂e, and subsequently offsetting the corresponding CO₂e by reforestation will not help to reach net-zero for these gasses. There is only one effective measure and that is keeping them out of the atmosphere.

Jevons paradox or the rebound effect

Jevons paradox states that more efficient use of resources will not only lead to a smaller ecological footprint, but also to a falling cost of products and services which might in turn lead to an increase in the demand, undoing the benefits of the efficiency gains. This is very noticeable in consumer goods. Due to globalization and technological improvements, consumer goods have become much cheaper than e.g. 40 years ago. In the assumption that the disposable income stays the same, consumption will increase, leading to a growing ecological footprint, and in some cases even bigger than the original one. Where in 1985 people could afford one basic desktop computer of 2000 euro, with the equivalent amount of 5700 euro in 2023, they can buy each member of a family of four a basic laptop, a tablet, a smartphone and a smart watch, inevitably leading to more emissions. Hence, efficiency is not a valid proxy for sustainability.

Another example is the increase in performance thanks to Moore's law. Instead of reducing the power consumption and the cost per chip in the new technology node, the industry has preferred to produce more powerful chips within the original power envelope and for about the same price. The extra compute power has been used by the software industry to add nice-to-have features that have subsequently increased the requirements for the hardware (more memory, more storage, more cores). In the last decade we have seen a transition from hard disks to SSDs, which consume less power, but have an embodied carbon footprint that is 8 times bigger than that of hard disks. Hence, Moore's law has substantially increased the carbon footprint of computing up to the point that most consumer devices are today too powerful for the workload they have to run. Similarly, supercomputers seldom have an efficiency of

more than 5% because the standard software stack adds a lot of overhead.

Enablement to the rescue

In its SMARTer 2030 report (Global e-sustainability initiative, 2020), the Global eSustainability Initiative states that ICT could save 12 GtCO_{2e} in 2030 in other major industries like manufacturing, agriculture, construction and transport. Compared to the estimated 1.25 GtCO_{2e} emitted by the ICT-industry itself, it would mean that ICT is net carbon negative and that an unconstrained growth of the ICT industry is actually good for the climate: for every additional kg of CO₂ emitted by the ICT industry, another 10 kg CO₂ will be saved in the rest of the economy, which means that a datacenter of a supercomputer virtually extracts carbon from the air.

This is obviously a flawed reasoning as it counts the emission reduction twice: once in the sector that realizes the reduction, and a second time in the ICT industry.

Furthermore, there is no evidence that enablement leads to carbon-negative emissions. Over the last 40 years, the ICT industry has grown and contributed to efficiency gains in all sectors of the economy, but this has not resulted in reduced carbon emissions at all (the proof is that the emissions are still increasing, despite the massive use of computing technology). The higher efficiency has led to more consumption and increased revenues, but also more pollution and faster depletion of natural resources. Historically, there is no proof that enablement works for reducing overall emissions because of the rebound effect in the sectors that implement digital technologies (Charlotte Freitag, 2021). Still, enablement is widely used to justify *digital exceptionalism*, which means that the computing industry can be exempted from carbon reduction (in comparison to e.g. the cement and steel industry), because they realize their reduction through enablement. A similar argument was used in the 20th century to exempt airlines from paying fuel taxes because air travel was good to better understand other cultures and thus for world peace.

Until the day that there is proof that enablement actively reduces emissions, the best way for the IT industry to contribute to the net-zero targets in 2070 is to become net-zero itself. If they would

lobby for an exception to become net-zero in 2070 based on the enablement they claim to create, they will first have to show that enablement exists for their activities, and how big it is. Without hard evidence, they are lobbying for digital exceptionalism.

Then the question is: how can the IT industry reach net-zero by 2070. We refer to the other papers in the sustainable section for inspiration.

ICT for green

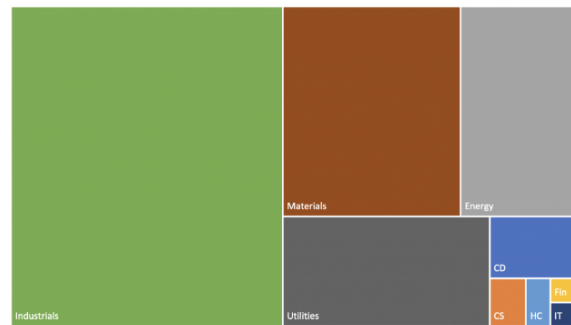
The European commission rightfully states that the green transition will require ICT to be successful. This is a no-brainer because optimization of resource consumption is at the basis of the green transition and advanced optimization requires computing. The statement that the green transition requires ICT is however not equivalent with the statement that ICT is good for sustainability (which is the assumption behind enablement). More concretely, it is not because some use cases in sectors like transport or energy can be made more sustainable with technologies like IoT, 5G or AI, that a widespread adoption of these technologies in the economy will have the same beneficial impact at a global scale. There is no proof of that.

Sustainability benefits of ICT will have to be proven use case per use case by comparing the scenario in which the technology is deployed, with the scenario in which the technology is not deployed or replaced by alternatives. The study should take into account the full life cycle, including all wanted and unwanted consequences of the technology used. An example is the self-driving car. The full optimization of the driving process will definitely have direct sustainability benefits: systematic eco-driving, congestion avoidance, less accidents. Combined with new business models like mobility as a service, it could also reduce the number of registered cars (which is good for the embodied energy, and requires less parking space, which could lead to more green spots in cities). Unwanted consequences could be that also children, very old people, and people without driving license can start using a car, leading to more car-km per year, and hence more energy consumption, or that people start working during their commute, and do no longer care about the duration of the trip (and hence about the jams). Before claiming that a self-driving car is more sustainable than a classical car, this analysis should be made.

Obviously, decisions are not only made based on sustainability criteria – there might be other criteria (safety, inclusion, increased productivity) that can justify the additional emissions. Based on such an analysis, it will be hard to prove that bitcoin calculations reduce the global carbon emissions (this seems obvious), or that the energy-hungry large language models will help reducing the global carbon emissions (this would be surprising, except for a few use cases). Other technologies like IoT and 5G might help reducing the carbon emissions in more use cases. Other cases are unclear: videoconferencing is always more sustainable than in-person meetings that require travel, but the convenience of videoconferencing has also replaced traditional phone calls and some intra-building business meetings leading to increased energy use (rebound effect).

Without additional legislation, a growth of the IT industry will not automatically lead to sustainability benefits. Sustainability benefits have to be proven case by case. Approaches like carbon accounting and Environmental, Social, Governance (ESG) reporting can help companies to understand their environmental impact, how to reduce it, and to prove that their IT solutions help in greening their company. It also guarantees that the reporting is standardized and can be used for benchmarking purposes. Eventually, this information could also be made available via the digital product passport (DPP) that might play a pivotal role in ending greenwashing practices within the industry. This information could further be translated into sustainability labels designed to assist consumers in making informed choices and increasing awareness about sustainability. These labels should be designed to be easily understandable and comparable to labels commonly used on food products or the euro-norms for car emissions.

According to ISS Insights (Aastha Agarwal, 2023), 57% of the European companies pledged net-zero commitment by 2050. In countries with a strong political net-zero commitment, more companies pledge to become net-zero by 2050, which proves that public policies help in making the industry more sustainable. Remarkably, the Information technology sector represents only a very small share of the companies that pledge net-zero (Figure 6).



Note: CD (Consumer Discretionary), CS (Consumer Staples), IT (Information Technology), Fin (Financials), HC (Health Care).
Source: ISS ESG

Figure 6. Companies with net-zero commitments, by sector. Source: (Aastha Agarwal, 2023)

According to IDC, the most common investments in sustainability are:

1. Investment into technologies empowering sustainable hybrid work environment (VPN, video conferencing, advanced security, etc.);
2. Use of recycled/recyclable materials in IT equipment/use of closed loop materials in IT;
3. IT infrastructure efficiency assessment and investments (energy-efficient devices);
4. Smart building energy efficiency related technologies and tools;
5. Carbon footprint/emissions tracking, data sourcing and data analysis;
6. Flexible consumption technology deployment models to eliminate resources overprovisioning.

The concrete order varies among sectors and among countries. According to the same study, the sustainability ambitions are greater than the sustainability actions. Only 50% of the companies have actually achieved the goals set forward for the last two years. There are multiple explanations: gap between strategic planning in the boardroom and execution on the work floor, difficulties implementing the ESG commitments, macroeconomic challenges (energy prices, inflation, political instability, ...), and the difficulty to present a clear business case with a reasonable return on investment for sustainability investments. In many companies, the sustainability transition requires a substantial change to stay in business. This is why sustainability is perceived as a risk management issue and compliance cost instead of business opportunity.

Proving the business case for sustainability is critical for widespread adoption, and ultimately the success of the European Union Green Deal. When sustainability is seen as a driver for business growth, it will be a door opener, getting traction from the top to the bottom of an organization. The EU or national governments could help by refocusing macroeconomic success away from GDP growth towards non-financial metrics to assess the well-being of the country.

Companies are struggling with preparing the workforce for a net-zero economy. Whilst it is true that the economic transition is creating tremendous new job opportunities, it also creates huge challenges in the short term. E.g., during the transition, transportation companies will have to service both an existing fuel-based fleet + a more recent electrical fleet (which requires a different skill set, different tools, often a different workshop).

Often forgotten: the importance of water

There is no electrical energy that is 100% emissions-free, but there are big differences between energy sources (Figure 7). Obviously, the use of fossil fuels causes most emissions, but even reservoir lakes have greenhouse gas emissions too due to decomposing organic material that ends up in the lake (e.g. methane).

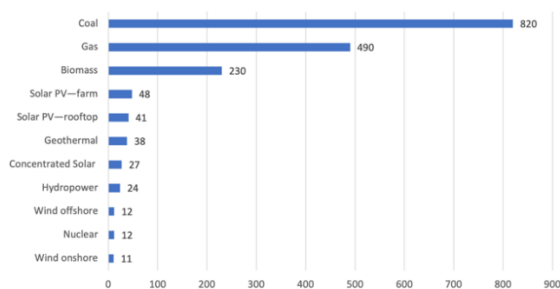


Figure 7. CO₂e emissions different electricity sources (in g per kWh)

In order to produce electricity, one does not only need fuel, but also water to (i) create the infrastructure (embodied water), and (ii) to close the thermodynamic cycle. Figure 8 gives an idea of the water footprint of the different types of energy generation. The Water Footprint is defined as the amount of water that is no longer available as drinking, irrigation, or process water after it has left the power plant. In most cases it is water that

evaporates from cooling towers, or from reservoir lakes.

A nuclear power plant of 1 GW produces around 4TJ of energy per hour and needs $610 \times 4 = 2440 \text{ m}^3$ of water per hour, which is about the capacity of an Olympic swimming pool per hour (2500 m^3). Coal and gas need $404 \times 4 = 1760 \text{ m}^3$ per hour. Firewood has a similar water footprint for the power plant but needs an enormous amount of water to grow the trees (Fuel supply). The high operational water consumption of hydropower might seem logical but the water that leaves the turbine is not ‘consumed’, it can be used downstream for irrigation, drinking or process water. The $15,100 \text{ m}^3$ is the amount of water per TJ that evaporates in the lake between the time it is collected (spring) and the time it is used to generate electricity. A lake evaporates around 1 m water per year (up to 2 m in deserts) which amounts to an average of $15,100 \text{ m}^2$ per TJ generated energy. The water footprint of solar is mostly determined by its embodied water. This is a consequence of the fact that one panel represents a lot of embodied water but produces during its lifetime only 5-10 MWh. A nuclear power plant produces the same amount of energy in 18-36s, and a windmill in 1h. The preparation of uranium, oil and coal requires lots of water too. The embodied water in the infrastructure and the fuel supply is extracted at the time and place of construction, the embodied water of the fuel is extracted near the mining site or in the refining factories. Only the operational water is spent at generation time and place.

One thing is clear: there is no electricity without water. Furthermore, the water consumption for electricity production is projected to double between 2010 and 2030 due to the increased electricity production, but also by investments in renewable energy sources like hydropower and biomass (firewood, energy crops, ...) that have a larger water footprint. Extended periods of drought might in the future limit the options to produce electricity in some parts of the world.

Data centers do not only need electricity to power them, but also need large amounts of water to cool them (and in the absence of water, they need electricity). In 2023 several data centers did not obtain a permit – because either the grid could not guarantee the power that was needed or because the water needed to cool the facility was not available. The water discussion is especially important in arid areas: Microsoft could not obtain a permit to

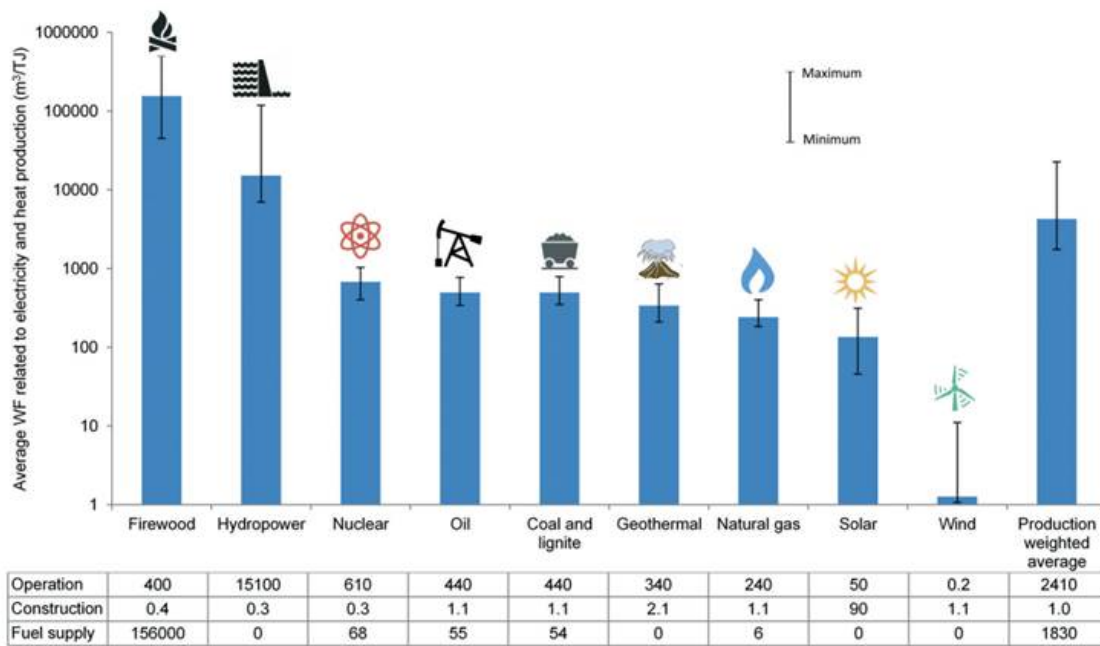


Fig. 2 Average consumptive WF per unit of electricity and heat produced ($\text{m}^3 \text{TJ}^{-1}$) for the period 2008–2012. Note that the scale is logarithmic. The ranges shown reflect minimum and maximum values per energy source. The values in the table represent the WF ($\text{m}^3 \text{TJ}^{-1}$) for the three main stages of the electricity and heat production chain.

Figure 8. Water footprint of electricity generation (Mesfin M. Mekonnen, 2015)

use water cooling in his new data center in Phoenix Arizona in the middle of the desert.

Not only data centers, but also fabs use lots of water. Today, a semiconductor fab uses around 8 l/cm² of ultrapure water (Wang, 2023), which means that a fab producing 20 000 wafers of 200 mm per month, needs more than one Olympic pool of process water per day. TSMC had to abandon a plan to build a 2 nm fab in a science park in the north of Taiwan after protest from the local residents about the space needed, and the amount of water it was going to consume (Mann, 2023).

Ireland is data center capital of Europe (O'Halloran, 2021) with 70 data centers in operation in 2021 (consuming 11% of the national electricity production). There will be more than 100 by 2025, and they are projected to consume 27% of the electricity production of the island. Adding one datacenter is like adding an extra town to the grid. Countries have to reduce their carbon emissions by 42% by 2030. It is clear that adding data centers is not helping them to reach their targets, and that governments prefer to invest in electricity generation for their citizens (to power their cars and heat pumps) rather than power data centers

from large hyperscalers. Similar discussions take place in The Netherlands and Singapore.

What do to with old devices?

The prevailing business models are rooted in the linear economic paradigm, which follows the take-make-consume-throw away cycle. Under this model, the more products a company produces and sells, the more successful it is perceived to be, benefiting the economy and society. This approach heavily relies on the availability of abundant and affordable raw materials and energy. However, as resources become scarcer, costs rise, and the environmental impact of waste becomes unsustainable (as exemplified by single-use plastics), there arises a pressing need to reduce the consumption of raw materials and energy. Additionally, waste should be viewed as a potential source of raw materials. Transitioning away from this linear model necessitates the development of new business models, based on a waste hierarchy. A waste hierarchy is a framework to reason about the end-of-life processing of devices. Europe has a Waste Framework Directive which is currently being revised. It consists of five levels (Figure 9).



Figure 9. Waste hierarchy

Prevention means that products are designed in ways to reduce waste: less pollution, less energy consumption, less harmful products or processes, less packaging, easier end-of-life processing. The European standardization of chargers is a good example of prevention.

Re-use can mean many things: devices can be sold as second-hand devices – even at a large scale by leasing companies; devices can be repaired – provided that they are repairable and spare parts are available; devices can be refurbished or remanufactured, i.e. updated to the current standards, or they can be repurposed, i.e., given a second life in a less demanding application context.

Recycling means that the materials of which the device is made are reinjected as raw materials in the industrial supply chains to create new products. Some fractions of e-waste are straightforward to recycle: glass, metals, plastics. Others like silver, gold, palladium and copper are more challenging but still economically viable. The semiconductors consist mostly of silicon with very small amounts of minerals deposited on it. Consequently, the silicon becomes 'contaminated' with these minerals, and their concentration is too low for economic (and ecological) extraction. Therefore, a die is an example of a non-circular product—it cannot be recycled at the end of its life. Hence, currently, the components that cause the majority of the emissions unfortunately cannot be recycled (which means that they cannot be part of the circular economy). This is not the only problem in this stage: globally, only a small fraction (17.4%) of the e-waste is currently being collected, and hence not recycled at all.

Recovery involves extracting energy from the fraction of the device that cannot be recycled (some plastics, glue, ...) by e.g. incinerating or gasifying it, and using the heat or gasses in a useful application. Incineration can be challenging if the

smoke contains toxic components that could pollute the environment, but it has the advantage that it also reduces the volume of the waste, and the ashes could be used in some applications (e.g. in construction).

Disposal involves storing the output of the recovery phase permanently into a landfill.

Currently, the world produces 53 million tonnes of electronic waste, or e-waste, and this will increase to 75 million tonnes by 2030. In 2019, only 17.4% of it was processed in recycling facilities. The remaining e-waste was either shipped to low or middle-income countries (where it often pollutes the environment), was thrown in the garbage bin, or kept at home for potential future use. In order to be sustainable, devices should be returned as soon as they are no longer used just like cars are returned when a new car is bought. This increases the chance on re-use and guarantees that the device will be properly recycled.

The most challenging level in the waste hierarchy, and the Achilles heel of the circular economy is recycling. Reuse can fairly easily be done by creating a market for second hand products (which is the norm for vehicles, houses, and other expensive goods), and recovery is standard in many western countries (incineration). Recycling is challenging however, not only because only a small percentage is being recycled, but also because the recycling technology for e-waste can only extract a limited number of raw materials from the e-waste. Ideally, these raw materials should be used to create new products. Unfortunately, we are far from this in the electronics industry. Very often the extracted materials do not have the same chemical properties to be used again in the electronics industry where ultrapure materials are required. They are therefore used in different sectors, just like recycled paper often ends up in cardboard. The fact that they are used in other industries is of course excellent and must be encouraged, but it makes the electronics industry non-circular because it remains dependent on freshly mined materials, some of which are in limited supply, and are used as geopolitical weapons by some countries. Hence, in order to become more sustainable, the electronics industry should invest in more effective recycling and try to become more circular.

Sustainability and time

Sustainability is now everywhere. As soon as an organization reduces its water or electricity consumption it calls itself sustainable. This is obviously incorrect. It is only sustainable if all the water and electricity consumption is sustainable. Even the expression that the organization has become 'more sustainable' is misleading as it suggests that it was already sustainable and now has become even 'more' sustainable. It would be better to state that the environmental impact or footprint has been reduced (which is very valuable in itself, but it is not synonymous with being sustainable, which can only be the ultimate end goal).

The ultimate goal for carbon emissions is to become carbon neutral by 2050 (and net-zero by 2070), and thereafter remain sustainable and net-zero forever. There is an intermediate checkpoint in 2030 to verify whether the joint global efforts are on track. Some people seem to believe that these deadlines are a kind of payment deadline by which a sum of money has to be paid back, and as soon as the debt has been paid, they can continue business as usual. This is obviously not the case. Now we have to work on a reduction of 43% by 2030, and after that we will have to work on the remaining 57% (which will probably be the hardest part).

Hence, becoming sustainable is a transition process in which the environmental footprint is gradually reduced until the moment that it is within the natural absorption capacity of the planet. For CO₂, this requires that we will have to stop emitting CO₂ from burning fossil fuels (either by stop using them, or capturing the emissions, and use them or store them). Hence, reducing the emissions of a process that emits 1000 kg CO₂ per day with 200 kg (per day) by adding a computing device with a carbon footprint of 20 kg per day is definitely good because it structurally reduces the total emissions from 1000 kg to 820 kg CO₂ from then on. The next question is how the remaining 820 kg can be further reduced to eventually be low enough that the remaining emissions are absorbed by natural processes. Hence, becoming sustainable is a continuous process.

Conclusion

The computing industry has an environmental footprint which will be growing during the next decade. In order to make the industry carbon

neutral, and eventually net-zero, more efforts will be needed. For the time being the two most important recommendations are to work on practical and universally accepted life cycle models for computing devices and services. Without such models, it is impossible to explore the design space for the most sustainable solution. In parallel, there is a need to educate the computing community on sustainability and to encourage it to start working on it and to make real progress.

Acknowledgement

The authors would like to thank Katharina Grimme and Angela Salmeron of IDC for their market insights.

References

- Aastha Agarwal, S. S. (2023, 5 16). Net Zero Pledges in Europe. Retrieved from ISS Insights: <https://insights.issgovernance.com/posts/net-zero-pledges-in-europe/>
- Charlotte Freitag, M. B.-L. (2021, September). The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations. *Patterns*, 1-18.
- D.H. Meadows et al. (1972, March 2). *The Limits to Growth*. Retrieved from <https://www.clubofrome.org/publication/the-limits-to-growth/>
- Global e-sustainability initiative. (2020). *ICT Solutions for 21st Century Challenges*. Global e-sustainability initiative (GeSI).
- Grober, U. (2007). Deep roots – A conceptual history of ‘sustainable development’ (Nachhaltigkeit). Retrieved November 27, 2022, from <https://bibliothek.wzb.eu/pdf/2007/p07-002.pdf>
- Mann, T. (2023, October 17). TSMC abandons plans for 2nm chip plant after Taiwanese locals protest. Retrieved from *The Register*: https://www.theregister.com/2023/10/17/tsmc_chip_factory_shelved/
- Mesfin M. Mekonnen, P. W.-L. (2015, 3 9). The consumptive water footprint of electricity and heat: a global assesment. *Environmental Science: Water Research & Technology*, 1, 285-297.
- O'Halloran, M. (2021, September 29). Ireland should pause development of data centres. *The Irish Times*.
- SEI, Climate Analytics, E3G, IISD, and UNEP. (2023). *The Production Gap: Phasing down or phasing up? Top fossil fuel producers plan even more extraction despite climate promises*. . Stockholm Environment Institute.
- The 17 Goals. (2015). (United Nations) Retrieved November 27, 2022, from <https://sdgs.un.org/goals>
- Triple bottom line. (2008, November 17). (*The Economist*) Retrieved November 26, 2022, from <https://www.economist.com/news/2009/11/17/triple-bottom-line>
- United Nations Environment Programme. (2023). *Emissions Gap Report 2023: Broken Record – Temperatures hit new highs, yet world fails to cut emissions (again)*. Nairobi. doi:<https://doi.org/10.59117/20.500.11822/43922>
- Vaclav Smil. (2022). *How the world really works*. Penguin books.
- Vincent Petit, S. C. (2021). *Digital economy and climate impact. A bottom-up forecast of the IT sector energy consumption and carbon footprint to 2030*. Schneider Electric Sustainability Institute.
- Wang, Q. (2023). Environmental data and facts in the semiconductor manufacturing industry: An unexpected high water and energy consumption situation. *Water Cycle*(4), 47-54.

Koen De Bosschere is a professor in the electronics department of Ghent University, Ghent, Belgium.

Patrick Blouet has been research and development program manager at STMicroelectronics with responsibility for European projects. For more than 25 years, he has been interested in environmental issues as a personal concern.



Information and communication technologies (ICT) are the fuel of the digital society. This article reviews the big challenges to make ICT (more) circular.

Sustainable materials and production

by Thomas Ernst, Lars-Åke Ragnarsson and Jean-Pierre Raskin

Europe is facing one of the largest challenges in its history: that of preserving a viable environment for the decades to come. On human consciousness of its environmental impact, Albert Allen Bartlett [1] said: “The greatest shortcoming of the human race is our inability to understand the exponential function”. ICT was, and still is, driven by exponential functions (Moore’s law, Cooper’s law, Koomey’s law, data exchange volume worldwide, ...). However, each exponential has its limits which are often reached sooner than expected as discussed further below.

Sustainability was defined by United Nations with a very broad vision, to promote prosperity while protecting the planet [2].

We will focus here on environmental aspects taken into account in the products Life Cycle Assessment (LCA) standards: resource consumption, emissions to the air, water, impacts on ecosystem, human health, and natural resources. The LCA goal is to guide product design and production to limit its impact. In ICT history to date, when some limits are reached, they are overcome by a technology breakthrough. By way of example, there was the CMOS (complementary metal-oxide-semiconductor) technology, a product of laboratory curiosity in the 1980’s when bipolar transistor power consumption was too high. New integrated systems (multi-core processors) and device (SOI, FinFET, ...) architectures entered into the market when energy consumption limits were reached by central processing units (CPUs). Today, CMOS is at the core of ICT and is the cheapest and most

reliably power-efficient technology for the development and growth of the internet of things (IoT).

Awareness of the rapid proliferation of wirelessly connected objects around us gives rise to questions about their environmental impact. There is a need to establish an ambitious roadmap, at research and industrial level, for moving into a virtuous cycle of eco-innovation based on precise environmental evaluations.

In this article, we propose pathways to maintain or even reduce global use of energy and critical materials for ICT.

Key insights

- ICT is energy and resource demanding and contributed to approximately 1,200 – 2,200 MtCO₂ eq in 2020, or equivalently 2.1%–3.9% of the global world greenhouse gas (GHG) emissions [3].
- Manufacturing of integrated circuits is greatly energy and material demanding and generates approximately 30% of the emissions of mobile ICT devices [3].
- The manufacturing of ICT uses many critical materials and Europe imports nearly all of them [4].
- Recycling of ICT is technically difficult, consumes energy, generates pollution, and is not currently profitable because of the low price

of primary raw materials. Only 17.4% of electronics waste is recycled [5].

- With an economic model more centred around service (functionality economy) than the object itself, reparability will be more encouraged.
- The “rebound effect” could often cancel out the efficiency gains of ICT systems.
- It is time to design differently, to design within limits.

Key recommendations

- Expand the lifetime of devices through more reliable design at the system level, both by enhancing the intrinsic durability of components (batteries, screen, memories, ...) and by adopting a modular approach in which the replacement of faulty or obsolete components is made easy.
- Rethink software and applications to make them less resource consuming (data, energy, hardware) and compatible with still usable hardware. Make software reconfigurable and evolutive.
- Integrate circular economy concepts, eco-design and full life cycle assessment (including fabrication, use and end-of-first-life phases) at the early stage of research and development of new ICT technologies.
- Promote research into alternative low global warming potential gases and chemical solutions. Abatement is a proven solution to reduce the impact of greenhouse gases used by the semiconductor manufacturers. Support its further adaptation, optimized maintenance and improvement.
- Avoid toxic materials where viable.
- As required by law, employ/assure that chemical processes are protective of human health and the environment.
- Develop manufacturing processes that facilitate materials recycling and low energy use.
- Develop products that are amenable to recycling at the end of life.
- Use secondary (recycled) materials, more renewable energy and develop bio-based and greener chemistry in the ICT industry. Develop and promote fabrication solutions

which use energy and materials more efficiently. This requires enhanced basic and applied research into materials and fabrication processes.

- Develop new economic models in which externalities of any product or service are properly accounted for and encouraging the manufacturer to make its best effort to decrease the environmental impact of its technologies. ICT solutions must be developed and implemented in such a way that the economic profitability is (at least partially) decoupled from the depletion of natural resources.

An increasing impact of ICT growth on natural resources

The digital society we live in has never been so energy- and material-intensive and this is leading to increasing pressure on natural resources, ecosystems and the climate. Today, ICT is a major economic activity and its impact is growing, similarly to transport, energy production, manufacturing, and agriculture. Information and communication technologies including TV's and other consumer electronics consume around 5% of the world's electricity production and were responsible for 2.1%–3.9% of the world greenhouse gas (GHG) emissions in 2020 [3], a level equivalent to air transport. Our smartphone contains electronic circuits that require more than sixty different materials. We are talking about virtually every element on Mendeleev's periodic table except radioactive materials. At the end of their life, the recycling rate of electronic equipment is very low (17.4%) [5].

It is extremely difficult to separate the sixty materials that make up electronic circuits [6]. Umicore, one of the most advanced companies in the field of electronic materials recycling, manages to extract 17 elements out of the 60 [6]. Material recovery from obsolete equipment is not profitable given the low cost of raw materials imported from the so-called global South. The environmental and social costs are not fully accounted for in the price of the final product. This leads to a double penalty for the global south countries: they suffer from environmental pollution (loss of biodiversity, pollution of the air and groundwater, etc.) during the extraction of raw materials, as well as receiving 75% of our electronic waste [7].

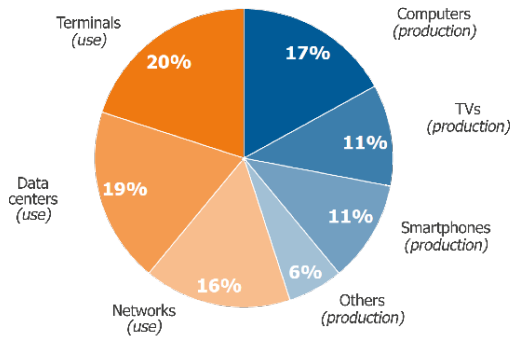


Figure 1 Estimation of distribution of the energy consumption of digital technologies for production (45%) and use (55%) Source: The Shift Project, 2019 [8]

Toward efficient integrated circuits production

The manufacturing of electronic devices used in the ICT industry is very complicated and requires large amounts of energy and materials (Figure 1). Especially the manufacturing of integrated circuits is an extremely complex and energy intensive endeavour done in very large fabs using hundreds of processing tools installed in clean room environment. IC manufacturing involves many types of complex process steps which use very high purity (also referred to as semiconductor grade) gases, chemicals, precursors, metals, and water. Some of the process steps use fluorinated gases which have very high global warming potential (orders of magnitude higher than CO₂) and therefore need to be abated to minimize the CO₂ equivalent emissions of a fab. The water used in wet cleaning processes needs to be ultrapure to minimize added particles and contaminants which could reduce yield and device performance. Finally, the production of the semiconductor grade materials required by the fab use complex and high energy demanding manufacturing processes themselves.

As a result, for mobile applications and devices, it is estimated that the manufacturing of IC's contributes significantly to the embodied emissions of a device. It is estimated that around 75% of the total Carbon footprint of mobile devices are from manufacturing, and about 50% of that (or ~40% of the total) is from the IC manufacturing of the many IC's [8].

The relentless efforts to develop denser memories and logic circuits continue to provide more functionality per chip area. However, this comes at a

cost in process complexity and an increase in fabrication steps. Furthermore, the new applications are becoming more sophisticated and require higher complexity at the system architecture level [9]. This results in an increase in the number of interconnect (metal) levels. As a result, more advanced semiconductor technologies have significantly higher embodied energy and associated emissions than former technologies for a chip with the same area. Generally speaking, newer technologies use more electrical energy, more raw materials, more chemicals and water. Figure 2 shows the evolution of the emissions related to process gases (Scope 1) and electricity (Scope 2) from the 28 nm to 5 nm estimated using a bottom-up virtual fab model [10]. The technology nodes are mapped approximately onto their year of introduction. For these technology nodes, the Scope 1 and 2 emissions, normalized per wafer, have increased by approximately 2x in the past 10 years.

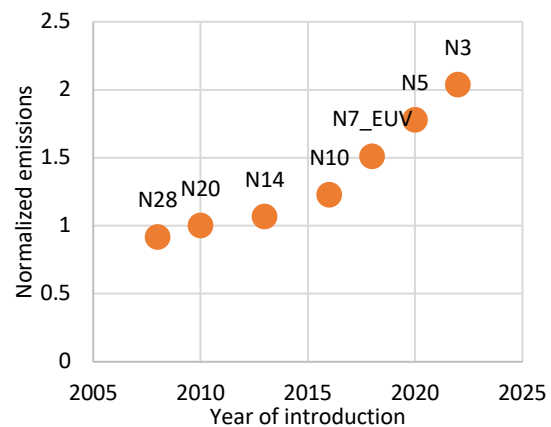


Figure 2 *imec.netzero v2.0.1 estimated emissions (Scopes 1 and 2) of imec process nodes representative of foundry nodes mapped onto their approximate year of introduction. The values are per wafer, normalized to the 20 nm technology node (N20) and use a 0.454 kgCO₂eq/kWh assumption for electricity [10]. The model is not taking fab nor electricity generation efficiency improvements over time nor electricity grid emission factor improvements over time into account.

Interesting to note is that when normalizing the emissions to the size of a function (e.g. a GB of memory, or a logic gate) the emissions are *reducing*. In other words, for a fixed function, a more advanced technology has lower emissions. However, due to the demand at the system level, this benefit rarely makes it to the final device. Instead, the improved scaling is used to create *ever higher complex circuits and systems* which end up *increasing* the total fabrication energy (rebound effect).

To enable these advances in technology scaling, inside the fab there are process steps that produce extremely high-quality materials and well-defined features with high aspect ratios and sub-nanometer precision. These processes require complex equipment using high purity process chemicals and gases. To keep the cost of manufacturing as low as possible, the efficiency of these processes is optimized as part of the technology development and maintenance. However, some processes are quite inefficient due to technical and process limitations. For example, the deposition efficiency using chemical vapor deposition (CVD) techniques of a material on the targeted wafer ranges from 1 to 20% [11]. This means that most (>80%) of the input high purity process gases is *passing through* the chamber to the exhaust without being actually deposited.

Furthermore, to enable the high density of advanced technologies there are many process steps which rely on the use of sacrificial layers. A first classical example is the photoresist used to transfer patterns from a mask (reticle) to the Si wafer surface. After development, the pattern is transferred to the surface by etching and then the resist is removed. Secondly, the deposition of “hard-masks” is introduced in many process flows nowadays to reach denser pattern densities when resist alone is not adequate. These hard-masks may include metals. Third, the definition of many three-dimensional structures (for example 3D NAND) relies on deposition of stacks of materials which are later etched and replaced by other materials through complex integration schemes. Finally, to enable high density lithography it is essential that the Si wafer remains flat throughout the manufacturing. This requires that material deposition filling small features will need to overflow and cover the full wafer and will be polished at the end using chemical mechanical polishing (CMP) resulting in further materials losses.

Due to these practices it is important to understand that the amount of material that ends up in the final chip is small compared with the amount of silicon, while the manufacturing processes require and waste significant amount of materials other than silicon. This puts some significant constraints on the materials circularity of IC manufacturing: It needs to happen inside the fab.

Efforts to reduce power consumption and industrial waste during the manufacturing of electronic devices and their components have been made

since the beginning of high-volume IC manufacturing. This work continues with for example, low temperature processes, reduction of heat dissipation in and from the oven, and reduction and recycling of chemicals and water. To take care of the direct emissions impact from fluorinated process gases, efficient abatement solutions have been introduced and are progressively used in the industry. There are also significant efforts towards finding replacement gases to lower the impact even further [12].

There are also good examples of circularity in the fab. Many manufacturers are today reusing their ultra-pure water [13] thus reducing their impact in regions where water scarcity is a concern. Other examples include recovery of copper from CMP slurry waste [14], recovery of H₂ used by extreme ultra-violet (EUV) scanners [15] and recirculation of hot water in cleaning processes [16].

Fab circularity and low-waste production

Beyond resource efficiency and recycling waste, the transition towards a cost-effective circular economy needs to be implemented as well as the general design methodologies of materials for sustainable development proposed by Ashby [17] adapted to specific ICT domains such microelectronics.

Some companies began to adopt the lifecycle analysis years ago [18], for example, in the production and recycling of the ultra-pure water needed [13] for the microelectronics industry. There are active research programs, both on the part of equipment makers and in research labs, seeking to reuse exhausted gases or fluids within the fab [19], and also to develop much more efficient material deposition techniques [11].

The reduction of toxic chemicals in the semiconductor manufacturing fabs is investigated through different methods. As an example, biowaste-based [20] chemicals and materials may be used to reduce significantly the use of solvents and chemicals in lithography, as suggested by preliminary results [21]; more research efforts are required. Bio-based materials are also being investigated for use in packaging in the ICT domain.

More circularity between companies means developing supply chains in which the by-products or

wastes from one industry could be the supply material for another. For instance, the hafnium required in CMOS production is a by-product of ultrapure zirconium used by the nuclear industry and produced in several countries including France and the United States [22].

To further reduce the waste of energy and materials, the ICT industry must adopt a holistic approach to developing sustainable products. Several initiatives already exist in the private sector. As an example, we can point out the significant and long-term efforts of several companies such as ST-Microelectronics which evaluates (by life cycle assessment) the carbon footprint of their microcontrollers [18] [23] and established a clear material declaration available online.

To envision a more sustainable future, Europe must:

- Take actions to make the ICT supply chain more transparent in collaboration with European Semiconductor Industry Association (ESIA);
- Systematically make life cycle assessment and declarations of materials (including for imported products) with shared methodology worldwide;
- Implement a clear and ambitious plan to maximize ICT product lifetime and anticipate its end of life.

This will encourage both research and industry sectors to innovate for the good of everyone.

Looking at minerals

The electronic industry needs a wide variety of minerals. For example, since the 1970's the silicon-based complementary metal-oxide-semiconductor (CMOS) field-effect transistor (FET) has been the mainstream technology for most transistor applications, thus making today's digital economy possible. Over the years, the number of elements used in their manufacture has increased greatly (Figure 3), especially since 2000 with the implementation of high-k dielectrics and metal gate stacks which are essential to minimize short-channel effects and gate-leakage current of short transistors (today gate length shorter than 20 nm).

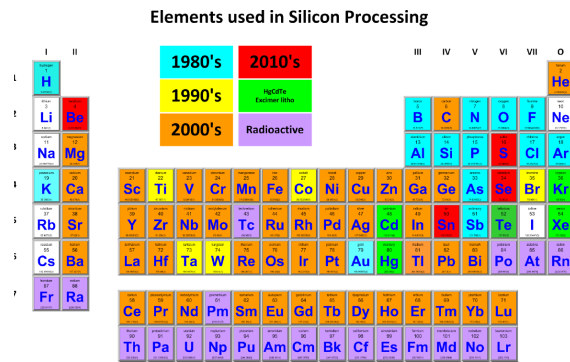


Figure 3 Introduction of elements in the manufacture of CMOS transistors: complementary metal-oxide-semiconductor (CMOS) transistors mainly involved silicon, oxygen, boron, phosphorus and integrated circuit interconnections were made of aluminum in the 1980s. There were relatively few changes in the 1990s, but a large diversity of elements was introduced in early 2000 and many integrated circuit interconnections were switched to copper [24].

A growing awareness of the limited nature of the supplies of some elements that have specialized and important uses is reflected in the proliferation of terms to describe them and the ores from which they are derived, including “gateway minerals” and “critical” or “endangered” elements. Some countries have adopted policies recognizing the high strategic importance of some of these for their physical and economic security.

In 2010, 14 elements were considered as critical by the European Commission (EC) according to both their strategic importance for future technology and their scarcity, while in 2023 the number rose to 34 [18].

Modern devices and systems rely heavily on a high degree of control of material properties and a mastery of manufacturing techniques and, to date, the ICT industry has been remarkably successful in fulfilling these needs.

The semiconductor manufacturing process is a “top-down” or “subtractive” one based on UV photolithography, etching and many sequential, highly organized and efficient steps of chemical and physical treatment of the chip, layer after layer.

Work is being undertaken on substituting or decreasing the use of hazardous and critical raw materials. For currently crucial elements such as indium, ruthenium, platinum, gallium, arsenic and gold, new technologies and materials are being investigated with a view to replacing them or drastically limiting their use in some critical devices (e.g. in sensors, memories, optoelectronics and spintronics) [25]. For instance, gold wire bonding has been replaced in IC packaging when possible

and replaced by other copper-based techniques like copper through silicon vias (TSV).

Other examples of the move towards sustainable electronics include the avoidance of lead in micro-components like actuators included in cell phones, use of 2D mono-atomic or ultra-thin atomic-deposition layers to reduce the use of some active materials by a factor of up to 10^6 , and use of silicon-based substrates such as silicon-on-insulator (SOI), instead of materials made from combinations of group III and group V elements, for radiofrequency (RF) technologies [26].

The scaling down of dimensions of high-tech devices in recent decades and the multiplication of materials in the components – some of them in extremely small quantities of a few micrograms – are leading to new challenges in recycling. The need for large amounts of power and the use of aggressive acids and solvents can make recycling of such electronics impractical [6].

Approaches to increasing the sustainability of microelectronic devices must include extending their lifetime through better design, by both enhancing the intrinsic durability and reliability of components and adopting a modular approach in which replacement of faulty or obsolete components is made easy.

These approaches can draw on the experience of, for example, some European microelectronics manufacturers and R&D laboratories (e.g. On-Semi, X-FAB, Infineon, STMicroelectronics, NXP, GlobalFoundries) that are designing or fabricating highly reliable components for automotive, energy management and security applications.

Steps towards the ambitious goal of achieving the sustainability of the physical layer of the digital society will require concerted actions covering a range of interlocked approaches. These will address the entire lifecycle of not only the digital devices themselves but also the services that support them, paying attention to their energy and environmental footprints as well as economy and efficiency in the utilization of resources.

Conclusion

The development of a more sustainable semiconductor industry is quite challenging and not only at the technical level. A mindset change has to be in-

duced not only in the industry but also at the research centre organizations and in the academia. Sustainability criteria must be defined at the early stage of development of any research project, product or service and their have to guide the choices of each stakeholder only the entire supply chain. There are encouraging signs such as the multiplication of policies at the European level to guide and support concrete actions for developing greener electronics, and strong engagement of several semiconductor companies to minimize their environmental footprint. In order to drastically reduce the ICT e-waste growth we have to multiply and disseminate the good practices in term of eco-design and circular economy.

Acknowledgement

The present article summarizes the main discussion outcomes from a panel of experts in the fields of geology, materials science, micro and nanoelectronics fabrication process, electronics circuits design, electronics packaging, supply chain management, and IT systems: Mathilde Billaud, Fraunhofer IZM, David Bol, UCLouvain, Thierry Baron, CNRS, Patrice Christmann, BRGM, Marie Garcia-Bardon, imec, Léa Di Cioccio, CEA-Leti, Laurent Pain, CEA-Leti, Bertrand Parvais, imec, Cédric Rolin, imec, Karine Samuel, UGA, Lutz Stobbe, Fraunhofer IZM.

References

- [1] A. Bartlett. [Online]. Available: <https://www.albartlett.org/>.
- [2] "www.un.org," [Online]. Available: <https://www.un.org/sustainabledevelopment/>.
- [3] C. Freitag, M. Berners-Lee, K. Widdicks, B. Knowles, G. S. Blair and A. Friday, "The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations," *Patterns*, vol. 2, no. 9, 2021.
- [4] "Critical raw materials - European Commission," [Online]. Available: https://single-market-economy.ec.europa.eu/sectors/raw-materials/areas-specific-interest/critical-raw-materials_en.
- [5] World Health Organization, "Electronic waste," 18 10 2023. [Online]. Available: [https://www.who.int/news-room/factsheets/detail/electronic-waste-\(e-waste\)](https://www.who.int/news-room/factsheets/detail/electronic-waste-(e-waste)). [Accessed 20 12 2023].
- [6] C. Hagelüken, "Recycling of Electronic Scrap at Umicore. Precious Metals Refining," *Acta Metallurgica Slovaca*, vol. 12, pp. 111-120, 2006.
- [7] A. Shamim, A. Mursheda and I. Rafiq, "E-Waste Trading Impact on Public Health and Ecosystem Services in Developing Countries," *International Journal of Waste Resources*, vol. 5:4, 2015.
- [8] U. Gupta, Y. G. Kim, S. Lee, J. Tse, H.-H. S. Lee, G.-Y. Wei, D. Brooks and C.-J. Wu, "Chasing Carbon: The Elusive Environmental Footprint of Computing," 2020. [Online]. Available: <https://doi.org/10.48550/arXiv.2011.02839>.
- [9] M. Koyanagi, "System Integration Technology Based on 3D Integration," in *IEEE International Meeting for Future of Electron Devices*, Kansai (IMFEDK), Kyoto, Japan, 2022.
- [10] "netzero.imec-int.com," imec, [Online]. Available: <https://netzero.imec-int.com>.
- [11] H. Pedersen, N. T. Barry and J. Sundqvist, "Green CVD—Toward a sustainable philosophy for thin film deposition by chemical vapor deposition," *J. Vac. Sci. Technol. A*, vol. 39, no. 5, p. 051001, 2021.
- [12] C. Jennings, P. Nguyen, S. Biltek and N. Stafford, "Advances in low GWP etch gasses," in *Proc. SPIE 12499, Advanced Etch Technology and Process Integration for Nanopatterning XII, 124990G*, 2023.
- [13] W. Den, C.-H. Chen and Y.-C. Luo, "Revisiting the water-use efficiency performance for microelectronics manufacturing facilities: Using Taiwan's Science Parks as a case study," *Water-Energy Nexus 1*, pp. 116-133, 2018.
- [14] X. Wang, G. Buer, W. Fan, L. Gao and M. Huo, "Copper removal from semiconductor CMP wastewater in the presence of nano-SiO₂ through biosorption," *Water Reuse*, vol. 11, no. 2, pp. 289-300, June 2021.
- [15] C. Bailey, "Hydrogen Recovery in EUV Lithography," *Semiconductor Digest*, Vols. 23-26, April/May 2023.
- [16] J. Wang, C.-T. . Ni, S. Ku and A. Tzeng, "TSMC Develops Hot DI Water Circulation System 2.0 for Wafer Cleaning Tool, Saving 380,000 kWh/yr of Electricity," TSMC, [Online]. Available: <https://esg.tsmc.com/en/update/greenManufacturing/caseStudy/67>.
- [17] M. F. Ashby, *Materials and Sustainable Development 2nd edition*, Butterworth-Heinemann, 2022.
- [18] "STMicroelectronics, "Footprint of a ultralowpower MCU," [Online]. Available: https://www.st.com/content/st_com/en/about/st_approach_to_sustainability/sustainability-priorities/sustainable-technology/eco-design/footprint-of-an-ultra-low-power-mcu.html.
- [19] V. Revankar and H. S. Lahoti, "CVD-semens monosilane reactor process with complete utilization of feed gases and total recycle", patent US 8,657,958 B2, 2014.". Patent US 8,657,958 B2, 2014.
- [20] M. Caillau, P. Crémillieu, E. Laurenceau, Y. Chevolut and J.-L. Leclercq, "Fifty nanometer lines patterned into silica using water developable chitosan bioresist and electron beam lithography," *J. Vac. Sci. Technol. B*, vol. 35, p. 06GE01, 2017.
- [21] I. Servin, A. Teolis, A. Bazin, P. Durin, O. Sysova, C. Gablin, B. Saudet, D. Leonard, O. Soppera, J.-L. Leclercq, Y. Chevolut, R. Tiron and T. Delair, "Water-soluble bio-sourced resists for DUV lithography in a 200/300 mm pilot line environment," *Micro and Nano Engineering*, vol. 19, 2023.
- [22] C. T. Matos et al., "Material System Analysis of Nine Raw Materials: Barytes, Bismuth, Hafnium, Helium, Natural Rubber, Phosphorus, Scandium, Tantalum and Vanadium," 2021. [Online]. Available: <https://doi.org/10.2760/677981>.
- [23] T. T. Pirson, T. P. Delhaye, A. G. Pip, G. Le Brun, J.-P. Raskin and D. Bol, "The environmental footprint of IC production: review, analysis, and lessons from historical trends," *IEEE Transactions on Semiconductor Manufacturing*, vol. 36, no. 1, pp. 56-67, Fe, vol. 36, no. 1, pp. 56-67, 2023.
- [24] S. Dhong and J. P. Colinge, "FinFET to Nanowire Transistor," in *DAC 2014 Tutorial*, 2014.
- [25] J.-P. Raskin, "No Digital Society without Sustainable Communication Technologies," in *Workshop Toward sustainable IoT from rare materials to big data at the 50th European Solid-State Device Conference – ESSDERC 2020*, Grenoble, France, 2020.
- [26] J. Raskin, "FinFET and UTBB for RF SOI communication systems," *Solid-State Electronics*, vol. 125, p. 73–81, 2016.
- [27] H. Ferrboeuf, "'Lean ICT towards digital sobriety' and 'Lean ICT Materials'," 2019. [Online]. Available: <https://theshiftproject.org/en/article/lean-ict-our-new-report/>.

Thomas Ernst is a chief scientist at CEA-Leti, France

Lars-Åke Ragnarsson is a program director at imec, Belgium

Jean-Pierre Raskin is a professor at the Université Catholique de Louvain, Belgium



Embodied emissions will soon be more significant than operational emissions.

Towards sustainable computer systems

by Lieven Eeckhout

Sustainability and climate change are a major challenge for our generation. Reducing the environmental footprint of computing implies that we consider the entire life cycle of computer systems including the embodied footprint (manufacturing and production) and operational footprint (device use). Analysing current trends reveals that the embodied footprint is, or will soon be, more significant than the operational footprint. To overcome the inherent data uncertainty regarding sustainability, this article advocates the use of a first-order model to assess the environmental impact of design decisions computer engineers and scientists can make based on first principles. This article further illustrates how this model enables making design trade-offs (in both hardware and software) to reduce the environmental impact of computer systems.

Key findings

- Improving computing-system sustainability is **more involved than minimizing carbon emissions during production and usage**. Material use (including rare-earth elements and/or minerals from politically unstable regions in the world) and ultra-pure water consumption are significant sustainability concerns related to chip production. **Even if all the energy consumed during production and use were green, the environmental impact of computing would still be significant, and growing.**
- The environmental footprint of computing continues to grow under current scaling

trends. When focusing on carbon emissions, embodied emissions are, or will soon become, the biggest contributor compared to operational emissions across the broad range of computing devices.

- **Embodied emissions are growing** at a fast pace because of **increasing demand for chips** and **increasing energy intensity of semiconductor manufacturing**. Perhaps contradictory to popular belief, improving the energy efficiency of computing systems does not necessarily make them more sustainable.
- Improving the energy and power efficiency of computing systems may lead to a rebound effect (Jevons paradox) which may be counter-productive to the environmental impact if the resulting **increase in demand outweighs the efficiency improvement**.
- Improving computing-system sustainability requires a holistic approach to computer architecture design and development, requiring multi-dimensional optimization including chip area, power, energy, performance.
- A first-order model enables making and assessing design trade-offs to reduce the environmental footprint of computing despite the inherent data uncertainty.
- **There is a role to be played for both hardware and software** to make computer systems more sustainable (or at least less unsustainable).

Key recommendations

- Computer architects should take a holistic approach when designing sustainable computer systems, and not solely focus on carbon emissions.
- Computer architects and engineers should primarily focus on **reducing the embodied footprint of computer systems**. Reducing the operational footprint is of secondary importance, although still significant.
- Decarbonizing the manufacturing process is not a panacea as it does not affect other sustainability concerns related to material use and extraction, chemicals and gases emitted, and ultra-pure water consumed during production.
- Computer scientists and engineers should be wary of **Jevons paradox**. **Efficiency improvements most often lead to a significant rebound effect**. Collaborating with entrepreneurs may yield new, **more sustainable business models** for computing.
- Computer architects should collaborate with various partners along the supply chain, user groups, and end-of-life recyclers to obtain **high-quality data to assess the environmental impact of raw material extraction, manufacturing, production, assembly, transportation, product use, maintenance, recycling**, etc.
- **Sustainability modeling tools** (both detailed models and high-abstraction analytical models) need to be developed, finetuned and validated to be able to **holistically balance the embodied and operational footprint** of computing devices.
- **Existing and emerging architecture paradigms** (multi-core processing, hardware specialization, core microarchitecture, speculation, chiplet-based integration, etc.) need to be **assessed and re-evaluated from a sustainability perspective**.

Sustainability versus climate change

Climate change is one of the grand challenges of our generation. The recent United Nations Climate Change report [1] in preparation for COP 28, the Dubai Climate Change Conference in November-December 2023, alerts that “*national climate action plans remain insufficient to limit global temperature*

rise to 1.5 degrees Celsius and meet the goals of the Paris Agreement”. While countries are making progress to trend down global greenhouse gas (GHG) emissions, current efforts are insufficient to limit global temperature rise to 1.5 °C by the end of the century. Much more effort is needed to keep this threshold within reach: “*greenhouse gas emissions need to be cut 43% by 2030, compared to 2019 levels*”.

Virtually all economic sectors contribute to global emissions. The five economic sectors that contribute most to GHG emissions are industry, electricity, agriculture, transportation and buildings, accounting for nearly 90% of emissions, according to the Organization for Economic Co-operation and Development (OECD) [2]. Freitag et al. [3] recently reported that information and communication technology (ICT) is estimated to contribute 2.1% to 3.9% of worldwide GHG emissions – on par with the aviation industry – and this contribution is rising. As computer scientists and engineers, it is our responsibility to limit ICT’s contribution to global warming, and, if possible, even decrease it.

While climate change is receiving increasingly wide attention – rightfully so! – it is important that we keep the broader picture in mind when reasoning about potential solutions. The broader picture relates to sustainability. The Brundtland report of the World Council on Economic Development from 1987 provides a broad, yet useful, definition of sustainability: “*Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs.*” This definition is an extremely powerful and unambiguous statement. It is a call for action for our generation: when generating economic activity and developing new devices and services, we should be wary of the impact this may have on future generations.

Sustainability hence relates to the efficient use of materials and energy, the environmental impact, water consumption, biodiversity, (socio-)economics, impact on human health, human rights, politics, policy, legislation, etc. Global warming is an important aspect of sustainable development, but it should not be the only focus. The extraction and use of raw materials and energy sources is another important aspect of sustainability. What to do when devices reach their end of life and how to repurpose e-waste is equally critical. Sustainability also affects and may require new legislation and business models to reduce pressure on raw-

material extraction and to incentivize a circular economy with a reduced environmental footprint.

Unfortunately, sustainable development in computing is often narrowed down to reducing energy consumption and/or transitioning towards green energy sources. However, making our computer systems more sustainable is much more involved than reducing energy consumption and/or using green energy. Carbon-free computing does not necessarily bring us to a more sustainable future. It is critically important to look at the bigger picture and incorporate the total environmental footprint. For example, Acun et al. [4] point out that a data centre that operates solely on renewable energy does not minimize the total carbon footprint because of the large number of solar panels, wind farms, and batteries needed to enable carbon-free operation. The reason is that the embodied carbon emissions to produce and manufacture the renewable-energy devices (solar panels, wind farms and batteries) outweigh the operational carbon emissions saved during the lifetime of the data centre. This implies that, to minimize the total carbon footprint of a data centre, a more holistic approach is needed that accounts for both the embodied and operational emissions, rather than just focusing on the operational side.

Rebound effects

A necessary condition to reduce the environmental footprint of computing is to make our devices more carbon-efficient, i.e., manufacturing and using a computing device should incur fewer GHG emissions. Unfortunately, while making devices more carbon-efficient is a necessary condition, it is not a sufficient condition because of potential rebound effects. For example, reducing the energy (or carbon) intensity of a device or service typically leads to a price reduction, which in turn stimulates consumption. If the increased consumption outweighs the energy intensity reduction, we end up with a net increase in environmental impact – exactly opposite of what we had envisioned! This is the well-known Jevons Paradox, named after William Stanley Jevons, who was the first to observe the rebound effect of the steam engine's improved coal efficiency leading to an overall increase in coal consumption [5]. Applied to computing, improving the energy or power efficiency of a computing device does not necessarily lead to a net reduction in environmental impact. Most often, an energy- or power-efficiency gain leads to increased usage and deployment, effectively increasing the

environmental impact of computing. For example, a more power-efficient server incentivizes data centre operators to host more servers within the data centre's power envelope, which (may) lead(s) to an increased embodied and operational footprint. Mitigating this rebound effect might need new business models and/or legislation and regulation to make sure that the reduction in per-device emissions leads to an overall decrease in emissions across all devices.

Understanding trends in environmental impact of computing

Before discussing what we, computer engineers and computer scientists, can do to reduce the environmental footprint of computing, it is important that we understand what the current trends look like. To do so, we make a distinction between embodied versus operational emissions [6]. (For the discussion that follows, we will mostly focus on GHG emissions, but several aspects also pertain to other environmental concerns, such as raw material extraction, water consumption, etc.)

Embodied emissions relate to raw-material extraction, manufacturing, assembly, transportation, repair, maintenance, and end-of-life processing. Operational emissions relate to product use during a device's lifetime. Embodied emissions can be further categorized in scope-1, scope-2, and scope-3. Scope-1 refers to the chemicals and gases used during manufacturing – this includes fluorinated greenhouse gases with orders of magnitude higher global warming potential than CO₂. Scope-2 refers to the energy consumption during chip manufacturing – this includes empowering the extensive production facilities with hundreds of manufacturing tools and requiring climate and humidity control. Scope-3 pertains to the energy consumption for the extraction and production of materials used for integrated circuit manufacturing.

Gupta et al. [6] performed a comprehensive survey of consumer devices from vendors including Apple, Google, Huawei, and Microsoft. They conclude that embodied emissions dominate for battery-operated devices such as wearables, smartphones, tablets, and laptops, while operational emissions dominate for always-connected devices such as speakers, desktop computers and gaming consoles. For data centres, most emissions are related to construction, infrastructure, and hardware manufacturing: interestingly, while total

energy usage is trending up – presumably because of increased server count and/or higher degree of consolidation (cf. Jevons paradox) – total operational emissions are decreasing for Facebook and Google, thanks to their policy of contracting and securing green energy sources to power their hyperscale data centres.

Eeckhout reformulated the well-known Kaya identity to project how the environmental footprint computing will change into the future [11]. The overall conclusion from this analysis is that the embodied emissions are continuing to grow under current scaling trends, and that embodied emissions already are, or will soon be, the biggest contributor. The fundamental reason is the increasing demand for chips (because of economic dynamics based on selling products, i.e., linear economy) and the growing energy intensity of semiconductor manufacturing (because of advancements in chip technology), which do not seem to be counterbalanced by the transition to green energy sources and improvements in per-device energy and power efficiency.

Inherent data uncertainty

A major challenge when doing research in sustainable computing is the high degree of uncertainty along a variety of dimensions. While companies' sustainability reports and product lifecycle-assessment (LCA) reports provide a wealth of data, there remain many unknowns and data limitations, in part because of industry secretiveness, or simply because of lack of reliable data. For example, a recent study by imec [7], which attempts to quantify the environmental footprint of modern-day chip manufacturing, makes assumptions regarding the energy consumption of a fab's facility equipment (i.e. it is *“assumed to contribute to 40% of the total energy”*); furthermore, the degree of abatement of fluorinated GHGs (scope-1) is unknown, as well as the use of materials and the energy needed for material extraction (scope-3). As another example, the Apple iPhone12 LCA report [8] uses industry averages when parameters are unknown for the production process, i.e. a company may not know the sustainability impact of its suppliers.

The operational footprint and its importance relative to the embodied footprint is even harder to assess, as it depends on typical user behaviour, product lifetime, and the geographic location of the user (which determines the carbon intensity of the user's power grid mix). Historical data could

be insightful, but it only provides a hint. Note further that product use may be subject to the infamous rebound effect, which may significantly shift the relative importance of the operational versus embodied footprint.

Overall, it is safe to conclude that there is inherent data uncertainty. Gupta et al. [9] recently proposed the ACT model to analyse a computer system's sustainability at design time. This model relies on detailed numbers from production processes in industry. This is an important step for our community at large (both in industry and academia). Nevertheless, the authors note that there is *“lack of up-to-date carbon emission data for the latest compute, memory, and storage technologies”*. Furthermore, they hope to *“encourage industry to publish more detailed carbon characterizations to standardize carbon footprint accounting”*. Imec's sustainable semiconductor technology and systems (SSTS) program aims at addressing exactly this issue by collaborating with major industry players to quantify the environmental impact of integrated circuit manufacturing [10].

Sustainable design based on first principles

And yet, despite the large degrees of uncertainty and the multi-faceted design problem, computer engineers and scientists need to make computer systems more sustainable. A potential solution is to revert to first principles and guide sustainable design decisions using a first-order model. First-order modelling should not be viewed as a replacement for, but rather as a useful complement to, detailed models like ACT and others. In fact, a detailed sustainability accounting method can provide initial data for a first-order model, and vice versa, a first-order model can provide directions where the detailed model should be further refined.

A first-order model uses proxies for the embodied and operational footprint that computer architects have control over, see for example [12] for more details about a first-order model for processor chips. A useful, first-order proxy for the embodied footprint of a chip is its die size, i.e. the larger the chip, the higher the embodied footprint for a given chip technology in terms of the energy and materials needed and the chemicals and gases emitted during production of the chip. A useful proxy for the operational footprint of a chip is energy consumption assuming a fixed-work scenario (i.e. a device performs a fixed amount of work during its

entire lifetime) and power consumption assuming a fixed-time scenario (i.e. a device is used for the same amount of time, and hence performs more work). The relative importance of embodied versus operational emissions can be captured via a parameter α which the architect can vary to explore different use case scenarios.

What can we do on the hardware side?

Although (deliberately) simple, a first-order sustainability model can reveal a variety of interesting insights which computer architects can take forward to design more sustainable computer systems despite the inherent data uncertainty. There is a fruitful avenue of future work to explore how computer architectures can be made more sustainable. We provide three examples here to illustrate the trade-offs one can make using the first-order model. Expanding and analysing to what extent a broader range of archetypal CPU and GPU design paradigms and solutions (e.g. caching, speculation, microarchitecture, acceleration, etc.) affect computer system sustainability would be extremely valuable.

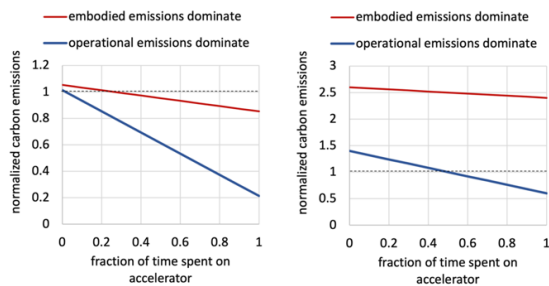


Figure 1: Total carbon footprint of a general-purpose CPU plus accelerator as a function of its degree of use, assuming that the accelerator takes up 6.5% extra chip area (left) versus 2x extra chip area (right), normalized to a general-purpose CPU without an accelerator. The accelerator is assumed to consume 500x less energy than the general-purpose CPU for performing the same work. Two scenarios are considered: embodied emissions account for 80% of total emissions versus 20% of total emissions. The larger the chip area of the accelerator, the more frequently it needs to be used and the higher the relative weight of the operational emissions need to be for the accelerator to be sustainable. Taken from [4].

Hardware specialisation and dark silicon

As reported in [12], the first-order model can be used to assess whether hardware specialization is sustainable. Integrating a hardware accelerator next to a general-purpose processor incurs a cost in terms of embodied footprint (because of a larger chip) which may be compensated for by the

reduced operational footprint (because of lower energy consumption when using the special-purpose accelerator rather than a general-purpose CPU). In other words, the reduced operational footprint amortizes the increased embodied footprint.

The question is where the tipping point is. The larger the accelerator, the more frequently the accelerator needs to be used and the higher the relative weight of the operational emissions needs to be for the accelerator design to be sustainable, as illustrated in Figure 1, if the accelerator is taking up significant chip area, and the embodied emissions dominate, the reduction in operational emissions does not compensate for the increased embodied emissions.

This suggests that the current trend towards large system-on-chip (SoC) designs with dozens of accelerators that occupy a significant fraction of the chip and that are not powered on all the time due to dark-silicon constraints, may not be a sustainable design paradigm. A more fruitful, sustainable design paradigm might be to consolidate accelerator designs to a common-denominator accelerator that can serve multiple critical applications while incurring less chip area, thereby reducing the embodied footprint at the expense of an increased operational footprint, with a net improvement in sustainability.

Core microarchitecture

A second example, also taken from [12], considers four microarchitectures: (1) a low-power in-order (InO) core; (2) a high-performance out-of-order (OoO) core; (3) a Forward Slice Core (FSC) [13], a complexity-effective core microarchitecture that aims for a level of performance that is comparable to OoO while incurring a small area and power overhead compared to InO; and (4) an OoO core enhanced with Precise Runahead Execution (PRE) [14], an efficient hardware data prefetching technique.

Figure 2 reports the total normalized carbon footprint for these microarchitectures as a function of performance considering different scenarios: fixed-work versus fixed-time and embodied versus operational emissions dominating. Ideally, a microarchitecture should be situated in the bottom right: high performance at low environmental footprint. Several interesting conclusions can be reached from this analysis. First, some microarchitectures are clearly better than others, possibly

under (a) specific scenario(s). For example, under a fixed-work scenario (subfigures a and b), FSC and PRE are clearly better design options than InO and OoO, respectively, because they achieve higher performance at a lower environmental footprint. Second, different microarchitectures offer different trade-offs. For example, while PRE yields higher performance than FSC, it also incurs a higher environmental footprint. Third, whether a microarchitecture incurs a lower environmental footprint may depend on the scenario. While PRE reduces the environmental footprint compared to OoO under a fixed-work scenario (subfigures a and b), it incurs a (much) higher footprint under a fixed-time scenario. This suggests that PRE is subject to a rebound effect: because PRE yields higher performance, it can perform more work in the same amount of time, which, because of its higher power consumption, leads to a higher operational footprint, and as a result a higher overall footprint.

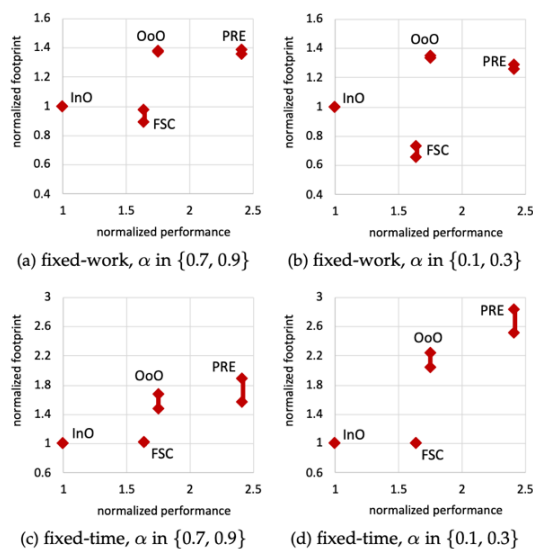


Figure 2: Comparing the OoO, Ino, FSC, and PRE microarchitectures in terms of normalised carbon footprint as a function of performance, assuming a fixed-work scenario (subfigures a and b) and fixed-time scenarioj (subfigures c and d) for different values for α , i.e. when the embodied footprint dominates (subfigures a and c) and when the operational footprint dominates (subfigures b and d). Taken from [12].

Chiplet-based integration

A third timely architecture trade-off worth exploring in the context of sustainability relates to chiplet-based integration. Small chiplets improve manufacturing yield, which reduces the amount of waste, and thus also the effective embodied footprint per correctly operating chiplet. Chiplet-based integration hence does not only reduce cost, but it also has the potential to improve sustainability [15]. Heterogeneous integration of chiplets

manufactured in different chip technology nodes could possibly further reduce the environmental footprint because older tech nodes incur a lower environmental footprint for the same chip area [7]. On the flip side, chiplet integration requires a silicon interposer or organic substrate with silicon bridges to connect the chiplets; these integration technologies and substrates obviously incur an additional environmental cost. When looking at the full picture, it is unclear whether homogeneous and/or heterogeneous chiplet integration reduces or increases the environmental footprint. Investigating these (and other) architecture trade-offs is a promising research avenue for computer engineers in industry and academia.

What can we do on the software side?

Reducing the environmental footprint of computing is not only a job for computer engineers – computer scientists can also contribute. In other words, this is not just a hardware problem, and software could be part of the solution. Of course, and most obviously, developers and researchers on the software side could (and should) aim for reducing the amount of energy and power consumed by software on existing (and future) hardware. This is a no-brainer, but it only affects the operational footprint of a computer system. There is an opportunity for software to also reduce the embodied footprint of computing. Two examples illustrate this.

Low-overhead programming languages

Pereira et al. [16] study the energy efficiency and memory consumption of a broad variety of programming languages. They conclude that high-abstraction, managed programming languages consume substantially more energy and memory than low-level, natively compiled programming languages. For example, they find that Python and Java consume 75x and 2x more energy, and 2.4x and 5.1x more memory compared to C, respectively. This implies that there is a direct reduction in operational footprint to be achieved by implementing software in native languages rather than managed languages. But there is also an indirect reduction to be exploited: software written in native languages can run as efficiently on less powerful hardware (with less compute and memory capacity). Because the less powerful hardware incurs a smaller embodied footprint, this could lead to overall reduction in environmental footprint. Of

course, there are many more design goals to consider than just performance and sustainability, including software productivity and security, but at least this example illustrates that there is a potential for reducing the environmental footprint of computing by reverting to low-overhead programming languages and/or by reducing the run-time overhead in (managed) programming languages.

Parallelisation

In a similar way, parallelising software has the potential to temper the need for ever more powerful multicore processors with increasing number of cores. A simple calculation using Amdahl's Law illustrates this. A multicore processor with 16 cores running software where 95% of the serial execution has been parallelised yields 17% higher performance compared to a multicore processor with 32 cores running software where 90% of the serial execution has been parallelised. The 32-core processor incurs a higher embodied footprint because the chip is (approximately) twice as big as the 16-core processor, and yet it achieves higher performance. The reason is that software is (slightly) more parallel. In other words, parallelising software is a more sustainable way to improve performance than increasing core count. Of course, parallelising software is challenging but, if successful, it can lead to an overall footprint reduction.

Conclusion

Improving computing-system sustainability is a challenging and multi-faceted problem. The embodied footprint is, or will soon be, a more important contributor than the operational footprint, primarily due to an increasing demand for chips and increased energy intensity of integrated circuit manufacturing. Decarbonizing the production process and use phase of compute devices is not a panacea, though, because it does not address other sustainability concerns including raw material extraction, chemicals and gases emitted, and ultra-pure water used during production.

What makes sustainable computer system design unique compared to traditional optimization criteria is that it requires a holistic approach considering chip area, power, energy, performance. The field of computer architecture specifically, and computer science and engineering in general, has only recently embarked on this endeavour.

Computer architects should continue to (1) collect high-quality data to assess the sustainability impact across the entire lifetime of a computing device, from raw-material extraction, transportation, manufacturing, assembly, use, repair, end-of-life processing, etc., (2) develop detailed and high-abstraction models to help designers evaluate the impact on sustainability at design time, and (3) analyse and revisit architecture design paradigms considering their sustainability impact. Overall, sustainable system design is an extremely timely and societally important topic where substantial innovation is to be achieved and expected in the following years.

References

- [1] “Climate Plans Remain Insufficient: More Ambitious Action Needed Now,” United Nations, 26 October 2022. [Online]. Available: <https://unfccc.int/news/climate-plans-remain-insufficient-more-ambitious-action-needed-now>. [Accessed 28 November 2022].
- [2] “Climate Action Explore policy solutions by key economic sector,” OECD, [Online]. Available: <https://www.oecd.org/stories/climate-action/key-sectors/>. [Accessed 28 November 2022].
- [3] C. Freitag, M. Berners-Lee, K. Widdicks, B. Knowles, G. S. Blair, A. Friday, “The Real Climate and Transformative Impact of ICT: A Critique of Estimates, Trends, and Regulations,” *Patterns*, vol. 2, no. 9, pp. 100340, <https://doi.org/10.1016/j.patter.2021.100340>, 2021.
- [4] B. Acun, B. Lee, K. Maeng, M. Chakkaravarthy, U. Gupta, D. Brooks, C.-J. Wu, “Carbon Explorer: A Holistic Approach for Designing Carbon-Aware Datacenters,” in *ACM International Conference on Architecture Support for Programming Languages and Operating Systems (ASPLOS)*, Vancouver, 2023.
- [5] “W. Stanley Jevons, “The Coal Question,” 1865,” Yale University, [Online]. Available: <https://energyhistory.yale.edu/library-item/w-stanley-jevons-coal-question-1865>. [Accessed 28 November 2022].
- [6] U. Gupta, Y. G. Kim, S. Lee, J. Tse, H.-H. S. Lee, G.-Y. Wei, D. Brooks, C.-J. Wu, “Chasing Carbon: The Elusive Environmental Footprint of Computing,” in *IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, Virtual, 2021.
- [7] M. Garcia Bardon, P. Wuytens, L.-A. Ragnarsson, G. Mirabelli, D. Jang, G. Willens, A. Mallik, S. Spessot, J. Ryckaert, B. Parvais, “DTCO including Sustainability: Power-Performance-Area-Cost-Environmental score (PPACE) Analysis for Logic Technologies,” in *2020 IEEE International Electron Devices Meeting (IEDM)*, Virtual, 2020.
- [8] “Product Environmental Report: iPhone 12,” 13 October 2020. [Online]. Available: https://www.apple.com/environment/pdf/products/iphone/iPhone_12_PER_Oct2020.pdf.
- [9] U. Gupta, M. Elgamal, G. Hills, G.-Y. Wei, H.-H. S. Lee, D. Brooks, C.-J. Wu, “ACT: Designing Sustainable Computer Systems with an Architectural Carbon Modeling Tool,” in *ISCA '22: Proceedings of the 49th Annual International Symposium on Computer Architecture*, New York, 2022.
- [10] L.-A. Ragnarsson, C. Rolin, S. Shamulila, E. Parton, “The green transition of the IC industry,” Imec, [Online]. Available: <https://www.imec-int.com/en/expertise/cmos-advanced/sustainable-semiconductor-technologies-and-systems-ssts/stss-white-paper>. [Accessed 28 November 2022].
- [11] L. Eeckhout, “Kaya for Computer Architects: Towards Sustainable Computer Systems,” *IEEE Micro*, pp. 1-8, <https://ieeexplore.ieee.org/document/9932869>, 2022.
- [12] L. Eeckhout, “A First-Order Model to Assess Computer Architecture Sustainability,” *IEEE Computer Architecture Letters*, vol. 21, no. 2, pp. 137-40, July-Dec 2022.
- [13] <https://dl.acm.org/doi/10.1145/3410463.3414629>
- [14] <https://ieeexplore.ieee.org/document/9065552>
- [15] <https://www.computer.org/csdl/journal/ca/2023/02/10244005/1QgWSmqbYt2>
- [16] <https://dl.acm.org/doi/10.1145/3136014.3136031>

Lieven Eeckhout is a senior full professor in the department of electronics and information systems at Ghent University, Belgium.

HiPEAC Vision 2024

The HiPEAC Vision 2024 recommendations focus on adapting European computing systems to rapidly evolving technologies and intense geopolitical and environmental pressures. Central to this is the development of the “next computing paradigm” (NCP), which integrates various key technologies – including artificial intelligence, (AI), the internet of things (IoT), and digital twins into a federated ecosystem. This paradigm emphasizes spatial computing, AI at the edge, dynamic web integration, and efficient orchestration of distributed services.

Advancing the NCP involves developing technologies essential for the NCP, making Europe a leader in AI, particularly for widespread edge use, innovating in hardware solutions, prioritizing cybersecurity, and promoting sustainability assessment for all systems. Global thinking and cross-domain collaborations are crucial.

Recommendations



1. Operationalize the NCP:

- Develop 4D-aware implementation technologies, standardizing representation and protocols for encoding physical objects and spaces, supporting mobile computation, and powering 4D-enabled operations.
- Augment APIs for interoperability, enhancing APIs with specifications for non-functional properties and dynamic service composition.
- Enable the mobility of computation, relocating data and processes as needed.
- Adopt (generative) AI at the edge for greater efficiency and privacy, and reduced latency.
- Develop AI-powered edge orchestrators that can dynamically combine services based on user needs.
- Encourage non-proprietary integration by supporting open standards and platforms.
- Initiate proof-of-concept efforts, creating demonstrators to showcase the advantages of NCP technologies.



2. Make Europe a leader in controlled AI development:

- Support EU growth in AI domains by investing in AI research and infrastructure.
- Develop foundation models based on “European” values and reflecting regional needs.
- Promote open-source AI models to enable access to shared AI resources.
- Develop local AI solutions and specialized accelerators for edge devices.
- Use AI for software and hardware development, upskilling engineers and researchers.
- Develop policies around AI accessibility and societal impact, ensuring equitable benefits from AI.
- Ensure “correctness by construction”, automating AI output verification for trustworthiness.



3. Innovations in hardware:

- Continue to improve performance and energy efficiency, including by exploring and integrating new technologies.
- Promote interdisciplinary research, driving new hardware paradigms.

- Explore innovative architectures for data-intensive computing.
- Develop a full European ecosystem, promoting fast prototyping and specialized architectures and developing the chiplet/interposer European ecosystem.
- Ensure sustainable hardware development, focusing on energy and emission reduction, and recyclability.



4. Cybersecurity as a core aspect:

- Address vulnerabilities, building systems with early consideration of cybersecurity and privacy.
- Reduce dependence on external ICT, using trusted European companies or open-source solutions instead.
- Address security challenges in large language models (LLMs).



5. Sustainability in computing:

- Create validated lifecycle models that model environmental impact.
- Develop sustainability-focused design, accounting for environmental costs in product design.
- Create viable sustainable business models for the ICT industry.
- Create ICT solutions for green applications in other industrial sectors.



6. Think globally:

- Promote collaboration: encourage teamwork across European research and technology groups.
- Cross-domain project calls: foster interdisciplinary research for innovative solutions.
- Competence centres: create centralized European expertise hubs.
- Build open-source ecosystems to accelerate innovation and accessibility.
- Multi-dimensional tooling: develop tools supporting capacity to address complex and critical challenges.
- Adopt a holistic approach to efficiency, emphasizing global co-design and system thinking.