

# Towards Trustworthy AI



G. Sharkov, AI HLEG member

+ some examples

# Background

## EU STRATEGY ON ARTIFICIAL INTELLIGENCE

published in April 2018

**Boost AI uptake**

**Tackle socio-economic changes**

**Ensure adequate ethical & legal framework**



In this context: appointment of Independent High-Level Expert Group on Artificial Intelligence (AI HLEG) in June 2018

# High-Level Expert Group and mandate

---

Chair:  
Pekka Ala-Pietilä

52 members from:



Industry



Academia



Civil society

## Two deliverables

- Ethics Guidelines for Artificial Intelligence
- Policy & Investment Recommendations

## Interaction with European AI Alliance

- Broad multi-stakeholder platform counting over 2800 members to discuss AI policy in Europe

# Ethics Guidelines for AI – Process



**18 December 2018**

First draft published

**December 2018-  
February 2018**

- Open consultation
- Discussion with Member States
- Discussion on the European AI Alliance

**March 2019**

Revised document delivered to the Commission

**April 2019**

Final document published & welcomed through Commission Communication



# Ethics Guidelines for AI – Intro



## Human-centric approach: AI as a means, not an end

Trustworthy AI as our foundational ambition, with three components

Lawful AI

Ethical AI

Robust AI

Three levels of abstraction

from principles  
(Chapter I)

to requirements  
(Chapter II)

to assessment list  
(Chapter III)

# Ethics Guidelines for AI – Principles

---

## 4 Ethical Principles based on fundamental rights



Respect for  
human  
autonomy



Prevention of  
harm



Fairness



Explicability

# Ethics Guidelines for AI – Requirements

---



Human agency and oversight



Diversity, non-discrimination and fairness



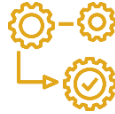
Technical Robustness and safety



Societal & environmental well-being



Privacy and data governance



Accountability



Transparency

To be continuously implemented & evaluated throughout AI system's life cycle

# Ethics Guidelines for AI – Assessment List

---



Assessment list to operationalise the requirements

- **Practical questions** for each requirement – 131 in total
- Test through piloting process to collect **feedback** from all stakeholders (public & private sector)

**Official launch of piloting: 26 June – Stakeholder event**



# Ethics Guidelines for AI – Piloting Process

---

- ❑ How to participate? Register today\*
  - Test out the assessment list
  - Provide us with feedback through an online survey
- ❑ In parallel: in-depth feedback process with selected stakeholders

\* <https://ec.europa.eu/futurium/en/register-piloting-process>



# Ethics Guidelines for AI – Best Practices

---

Fostering Best Practices on the Implementation of the Key Requirements

- Open page launched on the AI Alliance\*
- Collecting tools, methods, steps, other best practices to share with the community on how to achieve Trustworthy AI
- Everyone can contribute



\* <https://ec.europa.eu/futurium/en/ai-alliance-consultation/best-practices>



# Policy & Investment Recommendations

---



Second deliverable: different audience  
(Commission & Member States)

- Ensuring Europe's competitiveness and policies for Trustworthy AI
- Looking at key impacts and enablers
- Document to be presented at stakeholder event on 26 June 2019
- After the summer: recommendations for strategic sectors

# Scope: Policy & Investment Recommendations

## USING AI TO BUILD A POSITIVE IMPACT IN EUROPE

- Empowering and Protecting Human and Society
- Transforming Europe's Private Sector
- Catalysing Europe's Public Sector
- Ensuring World-Class Research Capabilities

## LEVERAGING EUROPE'S ENABLERS FOR AI

- Raising Funding and Investment for AI
- Building Data and Infrastructure for AI
- Generating appropriate Skills and Education for AI
- Establishing an appropriate governance framework for AI



# AI HLEG: Trustworthy AI

start: June 2018

In brief:

‘Trustworthy AI’ is the ‘**ideal**’ to which we aspire

- Trustworthy AI = (1) **Lawful AI** + (2) **Ethically Adherent AI** + (3) **Technically Robust AI**
- Each component is **necessary** but **not sufficient** to achieve Trustworthy AI.
- Ideally, all 3 components **work in harmony** and **overlap** in their operation.



# Trustworthy AI – the engineering perspective

Quality of AI =

**Quality of knowledge** + Quality of technology

+ Quality of software / hardware

+ (Cyber) security

*(+ the use in business models – ethical guidelines)*

AI systems & safety = “supervising” any ICT / SW systems (e.g. SCADA, ICS)

AI systems and autonomous defense/weapon systems =  
Explicable/Explainable AI

DARPA program – XAI (Explainable AI)

<https://www.darpa.mil/program/explainable-artificial-intelligence>

**DARPA** DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

MAIN MENU

Defense Advanced Research Projects Agency > Program Information

## Explainable Artificial Intelligence (XAI)

Mr. David Gunning

**AI System**

**DoD and non-DoD Applications**

- Transportation
- Security
- Medicine
- Finance
- Legal
- Military

**User**

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

• We are entering a new age of AI applications

• Machine learning is the core technology

• Machine learning models are opaque, non-intuitive, and difficult for people to understand



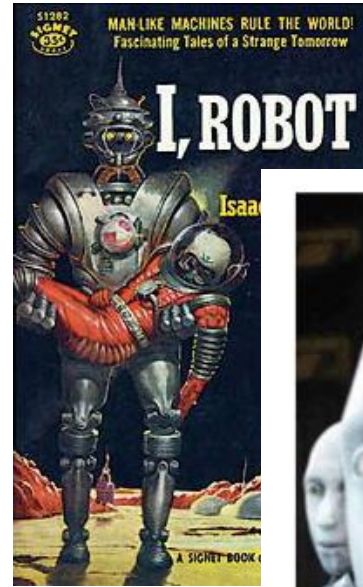
## Sci Fi or reality:

### The three laws of ROBOTICS

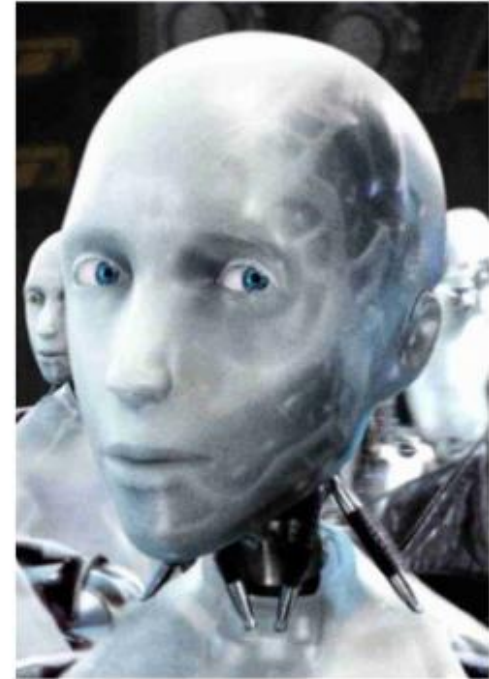
Isaac Asimov: 1942, story "Runaround"

1. A ROBOT may not injure a human being, or, through inaction, allow a human being to come to harm.
2. A ROBOT must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A ROBOT must protect its own existence as long as such protection does not conflict with the First or Second Law.

0. A robot may not harm humanity, or, by inaction, allow humanity to come to harm. (I. Asimov)
4. A robot must establish its identity as a robot in all cases. (L. Dilov)
5. A robot must know it is a robot. (N. Kesarovski)



When turning evil, display a red indicator light.



Tell me about this "friendship"  
thing you speak of...



**John McCarthy**, Stanford University  
Father of LISP language  
Introduced the term *artificial intelligence* in an August 1955

The long-term goal of AI is human-level AI.

I think the best hope for human-level AI is **logical AI**, based on the **formalizing of commonsense knowledge and reasoning in mathematical logic.**





# Examples: Chatbots or Intelligent Assistants? Public administration.



# AI vs. AI: Good Bots <> Bad Bots

## Good Bots

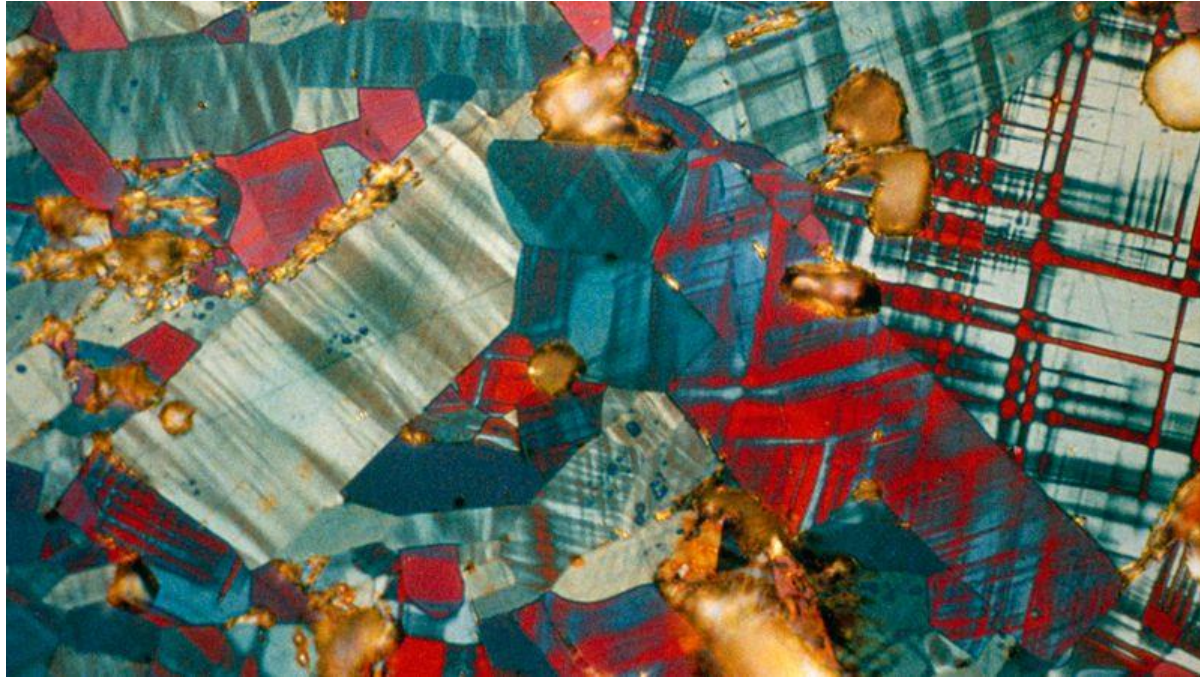
- Search Engine Crawling
- Website Health Monitoring
- Vulnerability Scanning



## Bad Bots

- DDoS
- Site Scraping
- Comment Spam
- SEO Spam
- Fraud
- Vulnerability scanning





“Using Machine Learning for Scientific Discovery in Electronic Quantum Matter Visualization Experiments” the team explores a 20 year-old hypothesis that could lead to the creation of a room-temperature superconductor. Team from Cornell, Harvard, Université Paris-Sud, Stanford, University of Tokyo and others





# Supporting SMEs – AI/ML as a service

**HACKERNOOIA** AI LATEST TOP 3.8 CRYPTO DEV POD AMA @ NOON W/ SECURITIZE CEO

## How Can AI Help Small Businesses?

Hey Siri! Can you help me with my business?

 **Mayank Pratap** [Follow](#)  
Feb 6 · 10 min read





**How AI Can Help Small Business and SMEs**

eb [www.engineerbabu.com](http://www.engineerbabu.com)

## Cheaper A.I. for Everyone Is the Promise With Intel and Facebook’s New Chip

Companies hoping to use artificial intelligence should benefit from more efficient chip designs

 **MIT Technology Review** [Follow](#)  
Jan 14 · 4 min read ★



A prototype of Intel’s NNP-I chip. Photo: Intel

<https://medium.com/mit-technology-review/cheaper-ai-for-everyone-is-the-promise-with-intel-and-facebooks-new-chip-497c34d591cb>



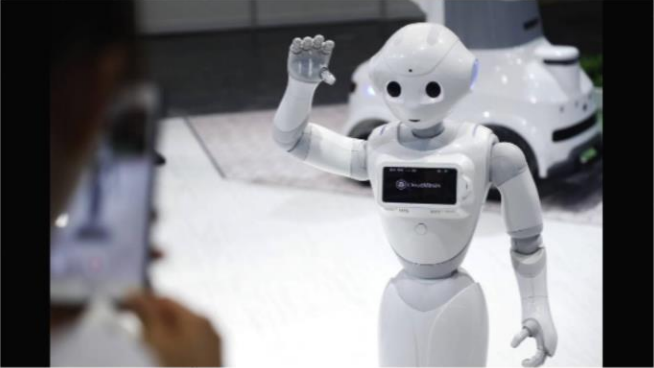
# AI hired, but new AI-related jobs..

WORLD NEWS • ARTIFICIAL INTELLIGENCE

## Artificial Intelligence Jobs Are on the Rise. Which Countries Are Hiring Right Now

Now

f t in m



Read More

By **RENAE REINTS** November 29, 2018

Artificial Intelligence (A.I.) is a massive industry, with potential in every field from [autonomous cars](#) to [human resources](#). According to PwC, A.I. could add up to [\\$15.7 trillion](#) to the global economy by 2030.

The [A.I. explosion](#) invites plenty of employment as well, although a [study](#) by Element AI found there's only about 90,000 people in the world with the right skill set.


FORTUNE

NEWS

BRIEFING • ARTIFICIAL INTELLIGENCE

## Finland Is Using Inmates to Help a Start-Up Train Its Artificial Intelligence Algorithms

f t in m





# DIGILIENCE 2019

Sofia, Bulgaria

2-4 October 2019

First International Scientific Conference "Digital Transformation, Cyber Security and Resilience" (DIGILIENCE 2019)

The rapid development and massive incorporation of advanced technologies transform industries, services, conflict, government, healthcare, leisure and social interaction. In the strive for competitive positioning, developers and users often underestimate safety and security considerations, which in turn provides ample opportunities for exploitation by malicious actors.

The series of DIGILIENCE conferences, the first of which will take place in the hearth of Sofia, the capital city of Bulgaria, aims to establish the state of the art and future demands in the provision of security and resilience of processes, services and systems that are heavily reliant on information technologies. Of particular interest are studies that examine systems in their interdependence or place their operation in a human or wider policy contexts, as well as evidence- and data-based studies and presentations of the respective datasets.

With these aims in mind, the Program Committee invites original contributions addressing the following themes:

- [Cyber Security Situational Awareness](#)
- [Detecting and Countering Malware](#)
- [AI for Cyber and Cyber for AI](#)
- [Intelligent Systems for Digital Forensics](#)
- [Fuzzy Methods for Cyber Security and Resilience](#)
- [Formal Methods and Model-based Security Testing](#)
- [Operations in Cyberspace](#)
- [The Human Factor in Cyber Security and Resilience](#)

[www.DIGILIENCE.org](http://www.DIGILIENCE.org)


2-4 October 2019



# QRS 2019 (IEEE) – Quality, Reliability and Security

## Sofia 22-26 July

### Workshop: CRE (Cyber Resilient Economy)



The screenshot shows the homepage of the QRS 2019 website. The browser address bar displays 'https://qrs19.techconf.org/'. The main header features the QRS 2019 logo, which consists of a stylized 'Q' in blue, red, and green, followed by the text 'QRS 2019 The 19th IEEE International Conference on Software Quality, Reliability, and Security July 22-26, 2019 • Sofia, Bulgaria'. Below the header is a navigation menu with links for Home, Registration, Submission, Tracks, Attending, Committees, and Series, along with a 'PC Login' link. The main content area is dominated by a large photograph of the Sofia University building, with the text 'Welcome to QRS 2019 Sofia University, Bulgaria' overlaid. Below the photo, there is a 'News & Announcements' section with three entries: 'Special Issue of Journal of Systems and Software' (dated 2018-10-10), 'PC Login accounts created' (dated 2018-08-10), and 'QRS 2019 CFP posted' (dated 2018-08-06). To the right of the news section, there are two highlighted boxes: 'Special Issue of JSS' and 'Journal First Papers'. The 'Special Issue of JSS' box contains text about inviting authors to submit extended versions to a special issue of the Journal of Systems and Software (JSS), with a 'Learn More' button. The 'Journal First Papers' box contains text about a partnership with IEEE Transactions on Reliability to include QRS papers.

**News & Announcements**

- 2018-10-10 **Special Issue of Journal of Systems and Software** [Download](#)
- 2018-08-10 **PC Login accounts created**
- 2018-08-06 **QRS 2019 CFP posted** [Download](#)

**About QRS**

In 2015, the SERE conference (IEEE International Conference on Software Security and Reliability) and the QSIC conference (IEEE International Conference on Quality Software) were combined into a single conference, QRS, with Q representing Quality, R for Reliability, and S for Security, sponsored by the IEEE

**Special Issue of JSS**

Authors of selected papers from QRS 2019 will be invited to submit an extended version to a special issue of Journal of Systems and Software (JSS).

[Learn More](#)

**Journal First Papers**

QRS has established a partnership with IEEE Transactions on Reliability to include



## Remember: Next steps



- 26 June: Presentation Recommendations & Kick-off Piloting
- Feedback gathering on assessment list from July till December 2019
- Revised version assessment list & sectorial recommendations in 2020
- Commission will then decide on Next Steps



# Thank you

