AI Safety 2019 workshop at IJCAI 2019

Towards

Trustworthy Autonomous and Intelligent Systems

Raja Chatila

Raja.Chatila@sorbonne-universite.fr

Institute of Intelligent Systems and Robotics (ISIR)

Sorbonne Université - Faculty of Science and Engineering, Paris

Chair, *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems*

Member the EU HLEG-AI

Member of CERNA – Commission de réflexion sur l'éthique de la recherche en sciences et technologies du numérique

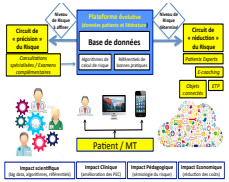
A Trustworthy **System**

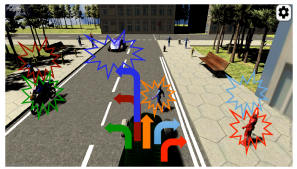
Decision Delegation to AI Based Systems

- AI is used in multiple applications sectors: transportation, health, justice, security, warfare, influence, insurance, finance, recruitment, management, personal service, assistance, …

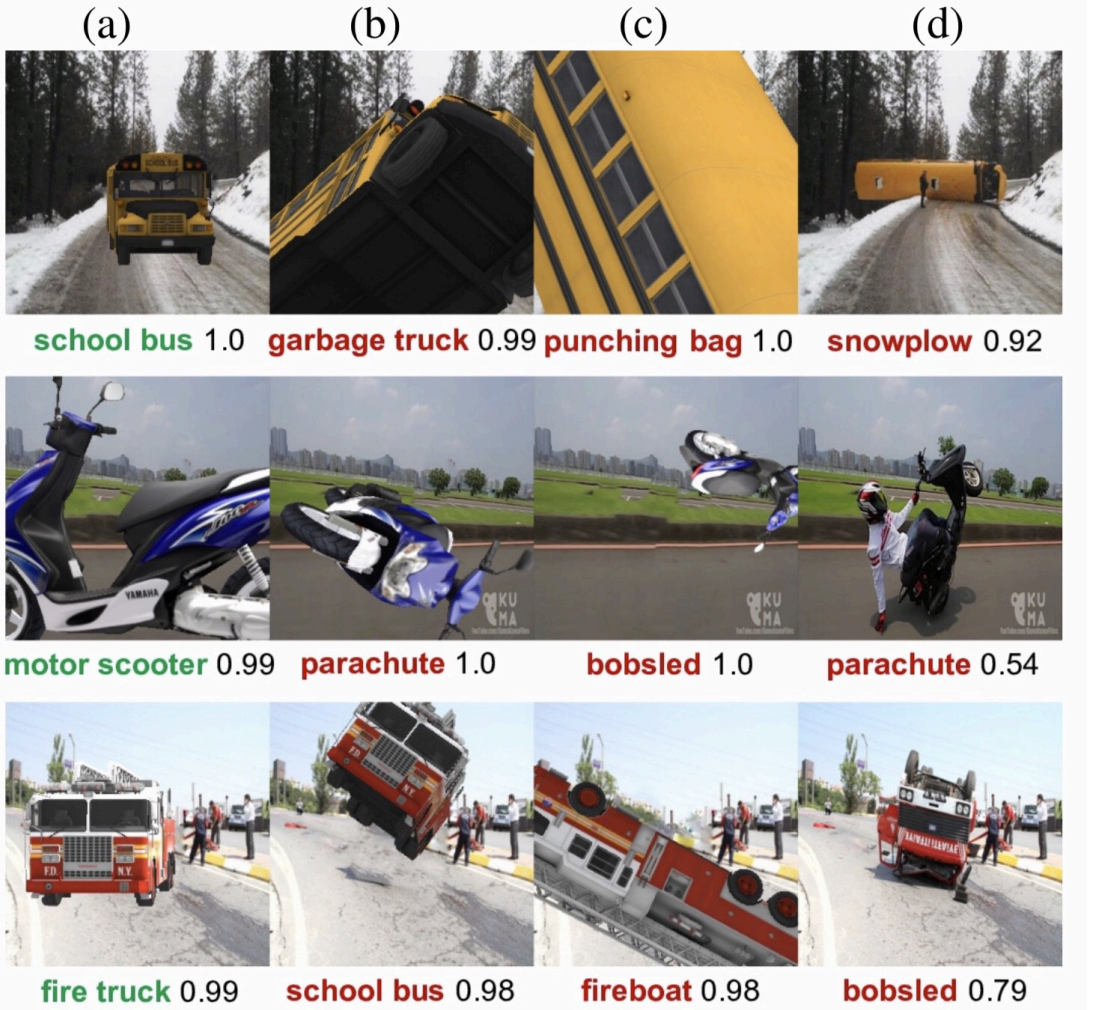
Issues:

- Critical applications (health, transport, HRI, …)
- Issues threatening human rights, wellbeing, fairness, … → **Ethics**
- Technical challenges for **reliability, safety, robustness**.
- Data driven ML not contextual, lack semantics



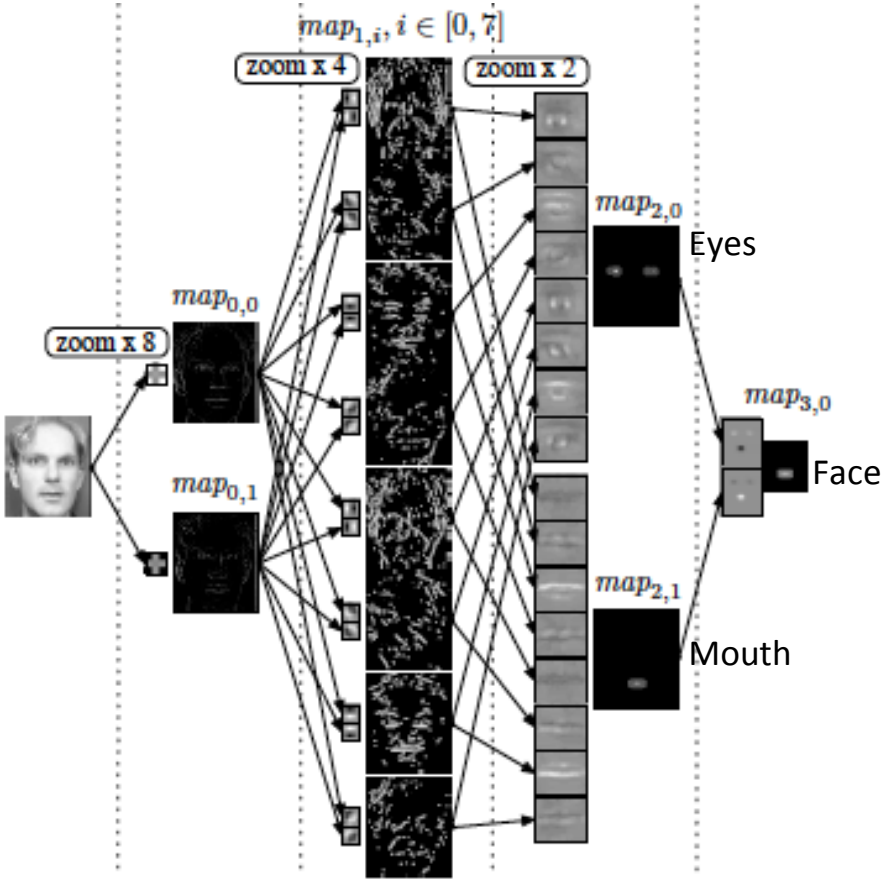


Limitations of Machine Learning Data Driven Interpretation



Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects. Michael A. Alcorn et al., April 2019

Bias and Transparency in Data and in Learning Processes



Bias issues

- Training data sampling the population
- Unbiased features
- Class semantics
- …

Architecture

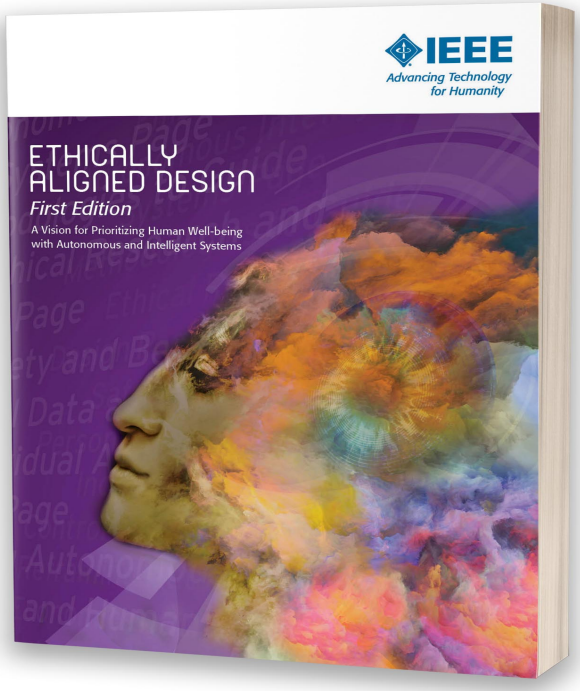
- Network structure and parameters

Trust in results

- Transparency: what is a “face”?

AI-Based Socio-Technical Systems

- Need to comply with human values
- Be technically dependable and socially trustworthy
- Need both **technica**l and **non-technical** frameworks



25 March 2019

8 April 2019

26 June 2019

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

Moving "From Principles to Practice" with standards projects, certification programs, and global consensus building to inspire the *Ethically Aligned Design* of autonomous and intelligent technologies

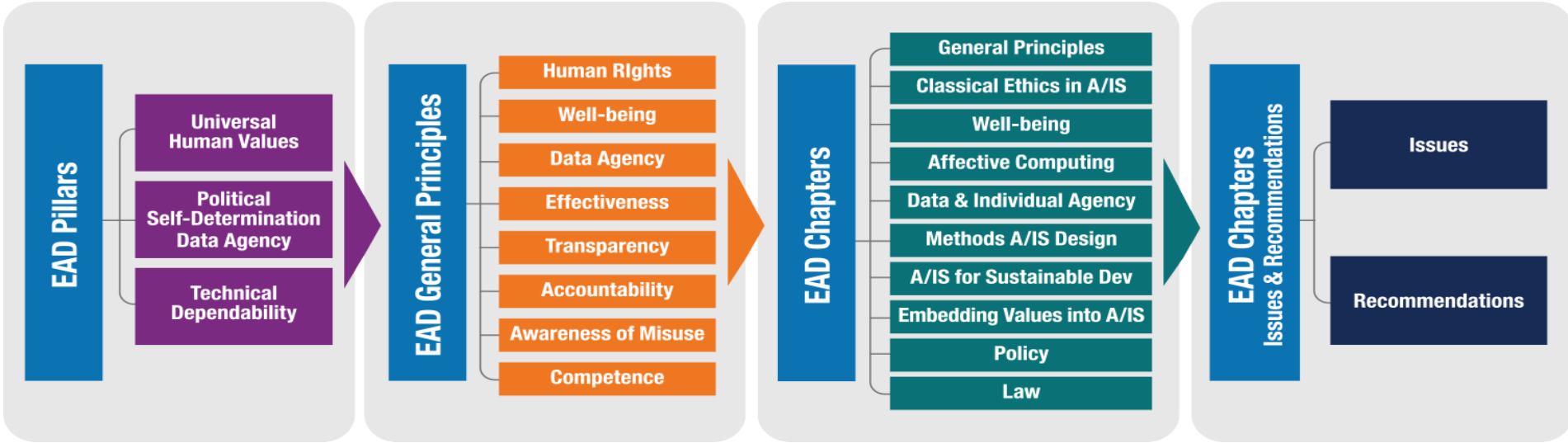
Mission: To ensure every stakeholder involved in the design and development of autonomous and intelligent systems is educated, trained, and empowered to prioritize ethical considerations so that these technologies are advanced for the benefit of humanity.

High-Level Community Statistics:

- Approximately 3000 individual members from all continents
- Strong regional groups (translating work into multiple languages)
- Around 40% women
- Rapidly increasing participation and endorsement by industry

IEEE

EAD1e - From Pillars to Practice



To download:
https://ethicsinaction.ieee.org

IEEE

EU HLEG-AI Ethics Guidelines Global Picture

Framework for Trustworthy AI

Trustworthy AI

- **Lawful AI**
- **Ethical AI**
- **Robust AI**

(not dealt with in this document)

Foundations of Trustworthy AI

Ensure adherence to ethical principles based on fundamental rights

→ **4 Ethical Principles**

Acknowledge and address tensions between them

- Respect for Human Autonomy
- Prevention of Harm
- Fairness
- Explicability

Realisation of Trustworthy AI

Ensure implementation of the key requirements

→ **7 Key Requirements**

Continuously evaluate and address these throughout the AI system's life cycle through

- Human Agency and Oversight
- Technical Robustness and Safety
- Privacy and Data Governance
- Transparency
- Diversity, Non-Discrimination and Fairness
- Societal and Environmental Wellbeing
- Accountability

→ **Technical Methods**

→ **Non-Technical Methods**

Assessment of Trustworthy AI

Ensure operationalisation of the key requirements

→ **Trustworthy AI Assessment**

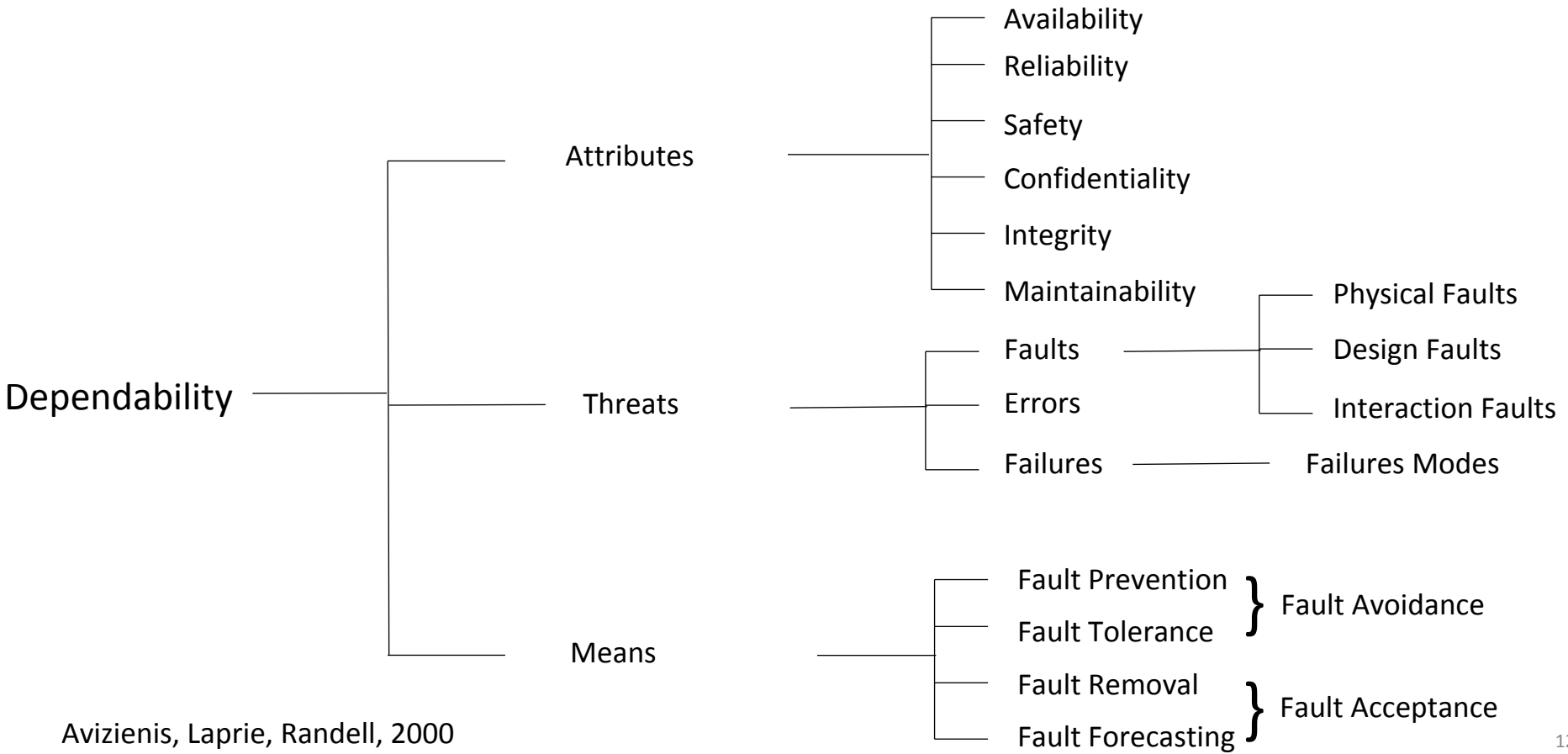
Tailor to the specific AI application

System Dependability

- Complex systems are built over interacting imperfect components (despite the best designs).

- System level – not component level - **Dependability** and **resilience.**

Dependability: Delivery of service that can <u>justifiably be trusted</u>
Resilience: The persistence of service delivery that can <u>justifiably be trusted</u>, when <u>facing changes</u>.



Avizienis, Laprie, Randell, 2000

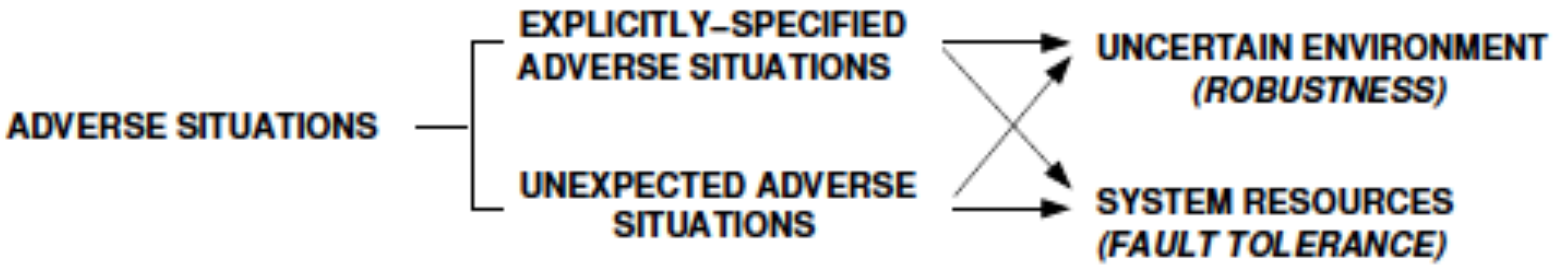
Dependability Attributes

- **Availability**: readiness for correct service;
- **Reliability**: continuity of correct service;
- **Safety**: absence of catastrophic consequences on the user(s) and the environment;
- **Confidentiality**: absence of unauthorized disclosure of information;
- **Integrity**: absence of improper system alterations;
- **Maintainability**: ability to undergo, modifications, and repairs.
- **Security**: availability for authorized users only + confidentiality + integrity (with ‘improper’ meaning ‘unauthorized’).

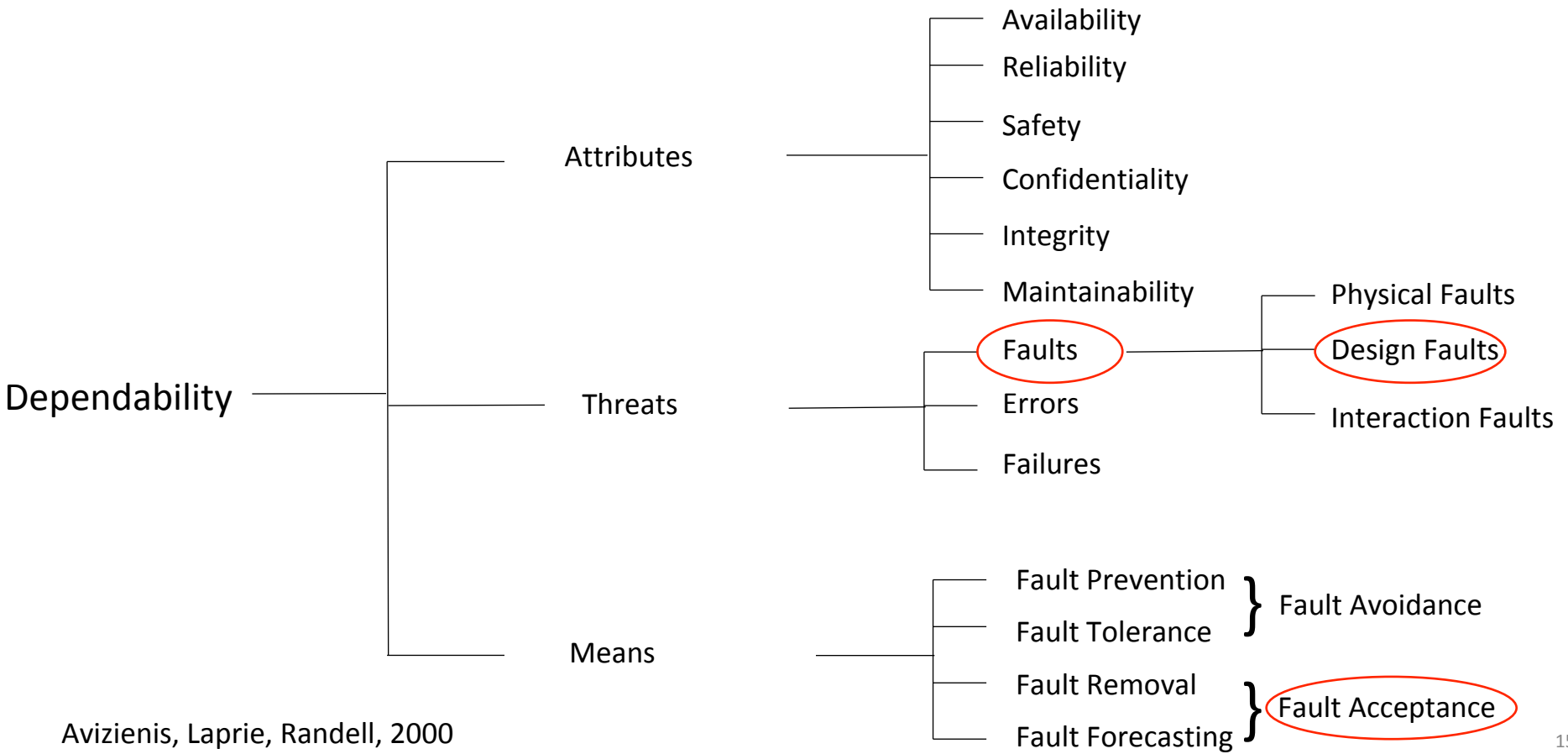
Dependability and Resilience

Autonomy and learning are mainly useful in open environments with uncertainties. But justifiably correct service cannot be guaranteed.

- **Fault Tolerance**: Ability to provide acceptable service despite system faults.
- **Robustness**: ability to provide acceptable service despite non explicitly specified environmental situations.



Dependability: Delivery of service that can justifiably be <u>trusted</u>
Resilience: The persistence of service delivery that can justifiably be trusted, when facing changes.



Avizienis, Laprie, Randell, 2000

Faults and Failures

- A system may **fail** either because it does not comply with the specification, or because the specification did not adequately describe its function.

- An **error** is that part of the system state that may cause a subsequent failure: a failure occurs when an error reaches the service interface and alters the service.

- A **fault** is the adjudged or hypothesized cause of an error. A fault is active when it produces an error; otherwise it is dormant.

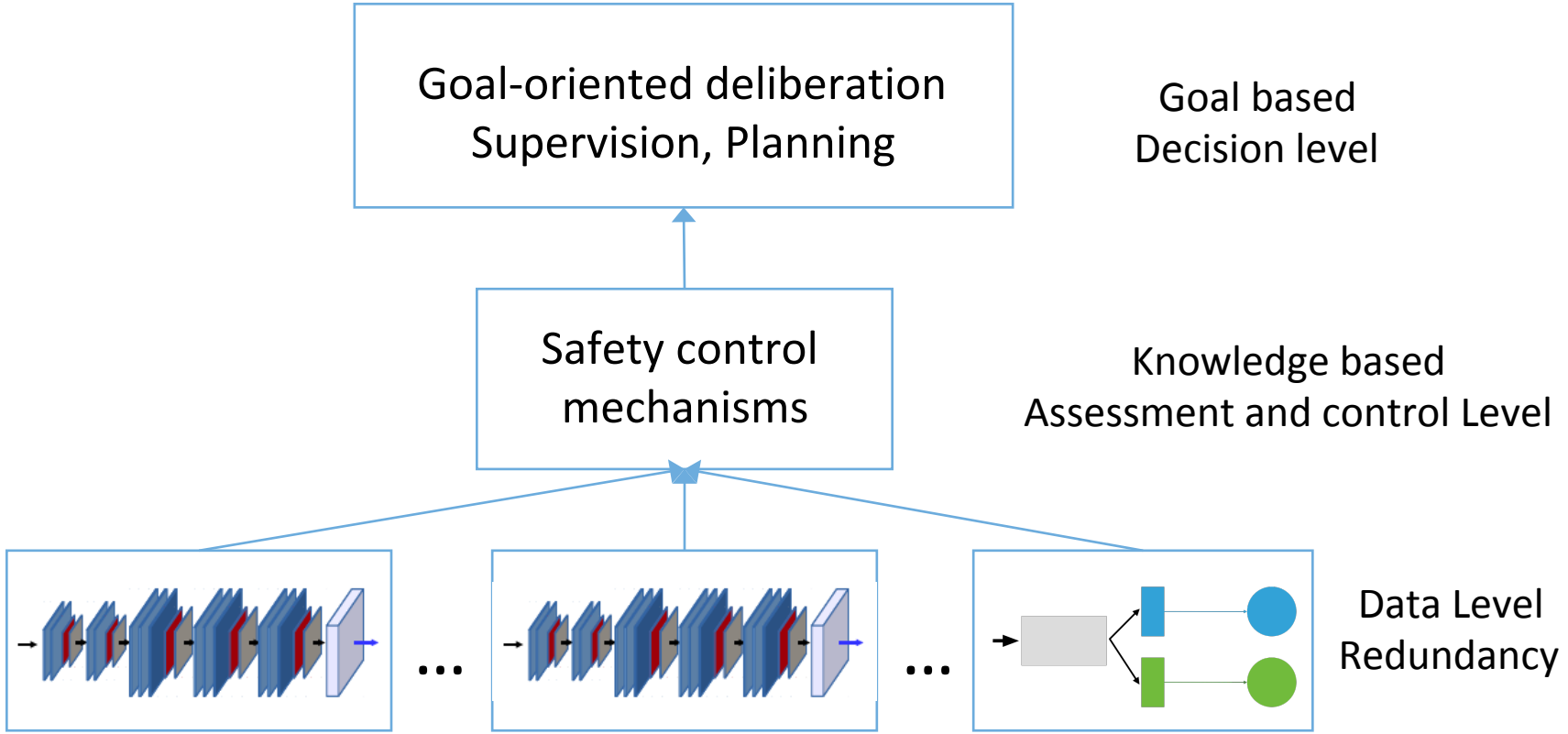
Avizienis, Laprie, Randell, Fundamental Concepts of Dependability, 2000

Fault Tolerance

- Limit consequences of task failure and maintain service continuity.
- Design diversity
- Detection of erroneous tasks to prevent propagation of errors: safety-bags, reasonableness checks; interception and rejection.
- Decision to produce error-free results

Lussier, Chatila, Ingrand, Killijian, Powell, 3rd iarp/ieee-ras/euron joint workshop on technical challenge for dependable robots in human environments, 2004

Fault Tolerance and Robustness System Architecture



Requirements for Trustworthy AI

1. **Human agency and oversight**- Including fundamental rights, human agency and human oversight
2. **Technical robustness and safety** - Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility
3. **Privacy and data governance** - Including respect for privacy, quality and integrity of data, and access to data
4. **Transparency** - Including traceability, explainability and communication
5. **Diversity, non-discrimination and fairness** - Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation
6. **Societal and environmental wellbeing** - Including sustainability and environmental friendliness, social impact, society and democracy
7. **Accountability** - Including auditability, minimization and reporting of negative impact, trade-offs and redress.

Achieving Trustworthy AI : Technical Aspects

- Architectures for Trustworthy AI
- Ethics and rule of law by design (X-by-design)
- Explanation methods
- Testing and validating
- Quality of Service Indicators

Trustworthy AI : Non-Technical aspects

- Regulation
- Codes of conduct
- Standardization
- Certification
- Accountability via governance frameworks
- Education and awareness to foster an ethical mind-set
- Stakeholder participation and social dialogue
- Diversity and inclusive design teams

IEEE P7000™ Standardization Projects

The IEEE P7000 series of standards projects under development represents a unique addition to the collection of over 1,900 global IEEE standards and projects. Whereas more traditional standards have a focus on technology interoperability, functionality, safety, and trade facilitation, the IEEE P7000 series addresses specific issues at the intersection of technological and ethical/societal considerations.

Like its technical standards counterparts, the IEEE P7000 series empowers innovation across borders and enables societal benefit.

For more information https://ethicsinaction.ieee.org/#set-the-standard

Standards Projects for Ethically Aligned Design

IEEE Advancing Technology for Humanity

IEEE P7000- Model Process for Addressing Ethical Concerns During System Design
IEEE P7001- Transparency of Autonomous System
IEEE P7002- Data Privacy Process
IEEE P7003- Algorithmic Bias Considerations
IEEE P7004- Standard on Child and Student Data Governance
IEEE P7005- Standard on Employer Data Governance
IEEE P7006- Standard on Personal Data AI Agent Working Group
IEEE P7007- Ontological Standard for Ethically driven Robotics and Automation Systems
IEEE P7008- Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems
IEEE P7009- Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems
IEEE P7010- Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems
IEEE P7011- Standard for the Process of Identifying and Rating the Trustworthiness of News Sources
IEEE P7012- Standard for Machine Readable Personal Privacy Terms
IEEE P7013- Inclusion and Application Standards for Automated Facial Analysis Technology
IEEE P7014- Standard for Ethical considerations in Emulated Empathy in Autonomous and Intelligent Systems

Realizing Trustworthy AI Over the System's Life Cycle