‘I trust you because you distrust AI’: auditing and zero-trust

Jerome De Cooman (jerome.decooman@uliege.be)

3rd October, 2025 Lawtomation Days

Technology and (dis)trust: AI between confidence and controversy

IE University, Madrid, Spain

What is trust?

- In a civil dispute over a contractual breach, both parties submit extensive written pleadings. To save time, the presiding judge uses ChatGPT to help summarise the arguments and identify key precedents. The judge verifies each citation, ensures that the reasoning aligns with the legal record, and ultimately issues a well-reasoned, accurate decision.
- When the judgment is published, it includes a short note disclosing that the judge used a generative AI tool to assist in drafting. Public debate ensues:
  - Was the judge's transparency commendable?
  - Should trust in the judiciary depend solely on outcome accuracy?

What is trust(worthiness)?

- Trust – risk – confidence (Luhman)
  - "Trust is an imperfect substitute for information" (Posner, 1978)
- Trust is relational
  - Trustor (subject of trust)
    - Lack of trust is "a main factor holding back a broader uptake of AI" (White Paper, 2020; Arrow, 1972)
    - Ergo: regulation is supposed to create an "ecosystem of trust" which in turn "should give citizens the confidence to take up AI applications" (White Paper, 2020)
  - Trustee (object of trust)
    - Interpersonal trust (I trust you)
    - Institutional trust (I trust the judicial power)
    - Technological trust (I trust the AI system used by the judge)

Distinction between trust and trustworthiness

- Laux (2023)
  - "Trust is improbable to be produced on demand (…) and impossible to achieve on command"
  - "Trustworthiness, on the other side, can be institutionally enforced, for example through contracts or audits with the threat of sanction"
  - "Enhancing trustworthiness can increase levels of trust if the increase in trustworthiness is recognized in the population"
    - Judges who use ChatGPT: transparency might be a way to enhance trustworthiness, but it seems it decrease litigants' trust in the judiciary

What is trust in the HLEG's view?

- "Trust is viewed as: (1) a set of specific beliefs dealing with benevolence, competence, integrity, and predictability (trusting beliefs); (2) the willingness of one party to depend on another in a risky situation (trusting intention); or (3) the combination of these elements." (Siau & Wang, 2028, quoted in HLEG, 2019)
- How can we achieve trust?
  - Enter trustworthiness

What is trust in the HLEG's view?

- Trustworthy AI has three component: it is (lawful), ethical and robust (HLEG)
- **Foundation of trustworthy AI:** adhere to ethical principles based on fundamental rights
- **Realisation of trustworthy AI:** implement the seven key requirements

Fundamental rights

1. Respect for human dignity
2. Freedom of the individual
3. Respect for democracy, justice and the rule of law
4. Equality, non-discrimination and solidarity
5. Citizens' right

Ethical principles

1. Respect for human autonomy
2. Prevention of harm
3. Fairness
4. Explicability

Key requirements

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and data governance
4. Transparency
5. Diversity, non-discrimination and fairness
6. Societal and environmental wellbeing
7. Accountability

What is trust in the HLEG's view? Focus on ethical principles

- "In situations in which no ethically ***acceptable*** trade-offs can be identified, **the development, deployment and use of the AI system should not proceed** in that form" (HLEG, 2019)
- "There may be situations, however, where no ethically acceptable trade-offs can be identified. **Certain fundamental rights and correlated principles are absolute and cannot be subject to a balancing exercise** (e.g. human dignity)." (HLEG, 2019)

What is trust in the HLEG's view? Focus on robustness

- "Trustworthy AI (…) should be robust, both from a technical and social perspective, since, even with good intentions, **AI systems can cause** unintentional **harm**" (HLEG, 2019)
- Robustness, as part of trustworthiness, means avoiding harms:
  - "Technical robustness requires that AI systems be developed with a preventative approach to risks and in a manner such that they reliably behave as intended while minimising unintentional and unexpected harm, and preventing ***unacceptable*** harm".

What is trust in the HLEG's view? Focus on robustness

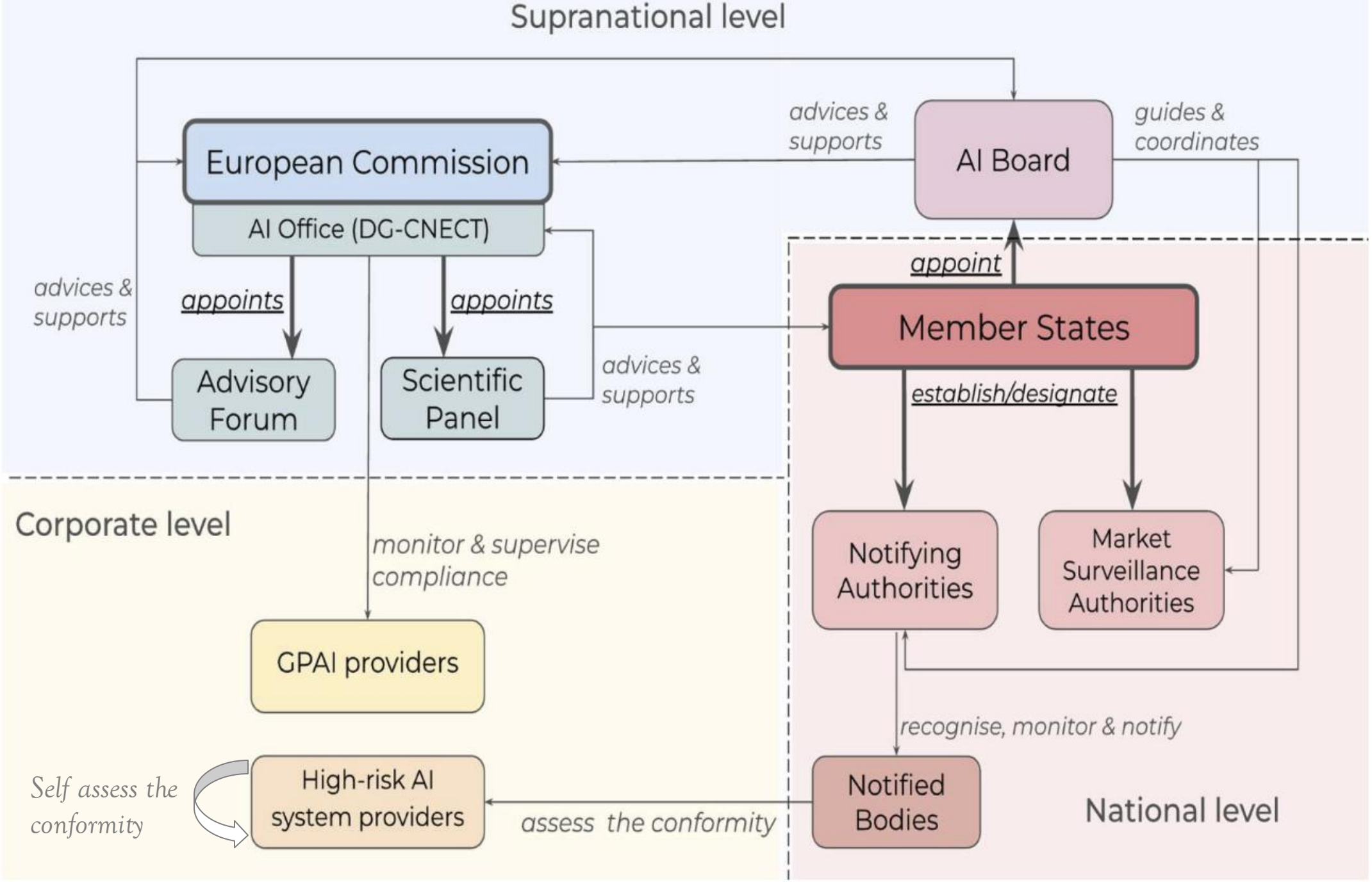
- Trustworthy AI Assessment List:
  - "Did you estimate the likely impact of a failure of your AI system when it provides wrong results, becomes unavailable, or provides societally ***unacceptable*** results (for example discrimination)?"
  - "Did you put in place ways to measure weather your system is making ***unacceptable*** amount of inaccurate predictions?"

What is trust in the HLEG's view?

- In short, under the HLEG's pen, trustworthiness is defined in light of the acceptability threshold
- Curiously, the EC's White Paper on AI does not discuss that threshold
- However, it reappears in the AIA
  - "a clearly defined risk-based approach (…) should tailor the type and content of such rules to the intensity and scope of the risks that AI systems can generate. It is therefore necessary to prohibit certain **unacceptable** AI practices, to lay down requirements for high-risk AI systems" (Recital 27 AIA)
  - "The risk management measures (…) shall be such that the relevant residual risk associated with each hazard, as well as the overall **residual risk of the high-risk AI systems is judged to be *acceptable***" (art. 9(5) AIA)

How to build trust?

- "The purpose of this Regulation is (...) to promote the uptake of human centric and trustworthy artificial intelligence" (recital 1; Art. 1 AIA)
- "In order to ensure a high level of trustworthiness of high-risk AI systems, those systems should be subject to a **conformity assessment prior to their placing on the market or putting into service**" (Recital 123 AIA)
- Whenever the AI Act grants discretion over what is acceptable, it demands a normative judgment (ie. Determining the acceptablity threshold) raising the key question of who should make that decision.



Adapted from C. Novelli et al., 'A Robust Governance for the AI Act: AI Office, AI Board, Scientific Panel, and National Authorities' (2024) *European Journal of Risk Regulation* 1-25

Trustworthiness in the AIA

- Decision No 768/2008/EC of the European Parliament and of the Council of 9 July 2008 on a common framework for the marketing of products (recital 22):
  - "The manufacturer, having detailed knowledge of the design and production process, is best placed to carry out the complete conformity assessment procedure"
  - "Conformity assessment should remain the obligation of the manufacturer alone."
- Providers of high-risk AI systems pursuant Annex III have to follow the self-assessment procedure (art. 43(2) AIA)
  - Exception: providers of biometric AI systems (Annex III(1) AIA) who has applied either harmonised standards (art. 40) or common specifications (art. 41) can choose between self-assessment (Annex VI AIA) or the assessment by a notified body (Annex VIII)
    - Upshot: the assessment by a notified body is mandatory for biometric AI systems if harmonised standards or common specifications do not exist or were not applied by the provider (art. 43(1) AIA)
  - The provider may choose any notified body, except if it is put into service
    - by law enforcement, immigration, or asylum authorities --> data protection supervisory authority
    - by EU institutions, bodies, offices and agencies → European Data Protection Supervisor

Increasing trustworthiness in the AIA

- According to Laux (2023):
  - One the one hand, "the reviewed normative literature appears to **prefer a more participatory model** of establishing trustworthiness in public sector AI systems (...). This creates some **tension with the expertocratic model of risk regulation pursued in the AI Act**"
  - On the other hand, "knowledge asymmetries can motivate an **additional requirement of trustworthiness, namely intermediaries which are themselves trustworthy**. The notions of 'trust proxies' or 'mediated trust' mentioned in the literature point in this direction"
    - Hence, the need for an increased role for notified bodies?

The role of notified bodies

- "Notified bodies shall be independent of the provider of a high-risk AI system in relation to which they perform conformity assessment activities. Notified bodies shall also be independent of any other operator having an economic interest in high-risk AI systems assessed, as well as of any competitors of the provider." (art. 31(4) AIA)
- Independence is a way to increase trustworthiness, but (Laux 2021):
  - Notified bodies provide their assessment by charging a fee;
    - They have an economic incentive to be required by a provider to assess the AI system
  - AI providers are free to choose any notified body;
    - There is a risk that notified bodies may prioritize AI providers' interests to secure repeat business.

The role of notified bodies

- Are third-party conformity assessment more efficient than self-assessment?
- On the one hand:
  - Research shows products certified by their own manufacturers (rather than by independent third parties) often do not actually meet the standards they claim to meet (Larson & Jordan, 2018)
    - toy safety recalls in the EU (conformity self-assessment) were 9 to 20 times greater than those in the US (third-party conformity assessment).
- On the other hand:
  - EU law requires medical grade silicon for breast implant
  - Breast implant (industrial silicon) CE certified by German notified body
  - 40,000 affected women in France; 400,000 worldwide

Failure of the model

- “In short, the New Approach has succeeded in fostering flourishing markets for certification services – but evidence suggests that it cannot be relied on systematically to deliver trustworthy products and services that protect individuals from harm to their health and safety” (Smuha and Yeung, 2024)

Conclusion & Recommendations

- Citizens trust is essential for AI's uptake in Europe
- *Trust* cannot be commanded ; *trustworthiness* can be institutionally enforced
- "Enhancing trustworthiness can increase levels of trust if the increase in trustworthiness is recognized in the population"
- The AIA aims at implementing the White Paper's ecosystem of trust
  - It is the regulatory framework necessary for the development of trustworthy AI

Conclusion & Recommendation

- The AIA conflates trustworthiness with risk acceptability
- Acceptability threshold is set by AI developers or, in rare case, notified bodies
- Studies show that, in light of asymmetrical information, the intermediary (expert who has the knowledge) must be trustworthy to increase trust level
- Independence is a way to increase trustworthiness
  - Although there is a risk of capture
- Although not a panacea, notified bodies appear to be more efficient than self-certification

Conclusion & Recommendations

- "Given the current experience of professional pre-market certifiers in the field of product safety and the different nature of risks involved, **it is appropriate to limit**, ***at least in an initial phase of application of this Regulation***, **the scope of application of third-party conformity assessment for high-risk AI systems other than those related to products**. Therefore, the conformity assessment of such systems should be carried out as a general rule by the provider under its own responsibility, with the only exception of AI systems intended to be used for biometrics." (recital 125 AIA)