The Explanations One Needs for the Explanations One Gives: Thoughts on the Epistemic Link between Explainable AI and Causal (Evidentiary) Explanations under the EU's AI Liability Regulation

LJUPCHO GROZDANOVSKI

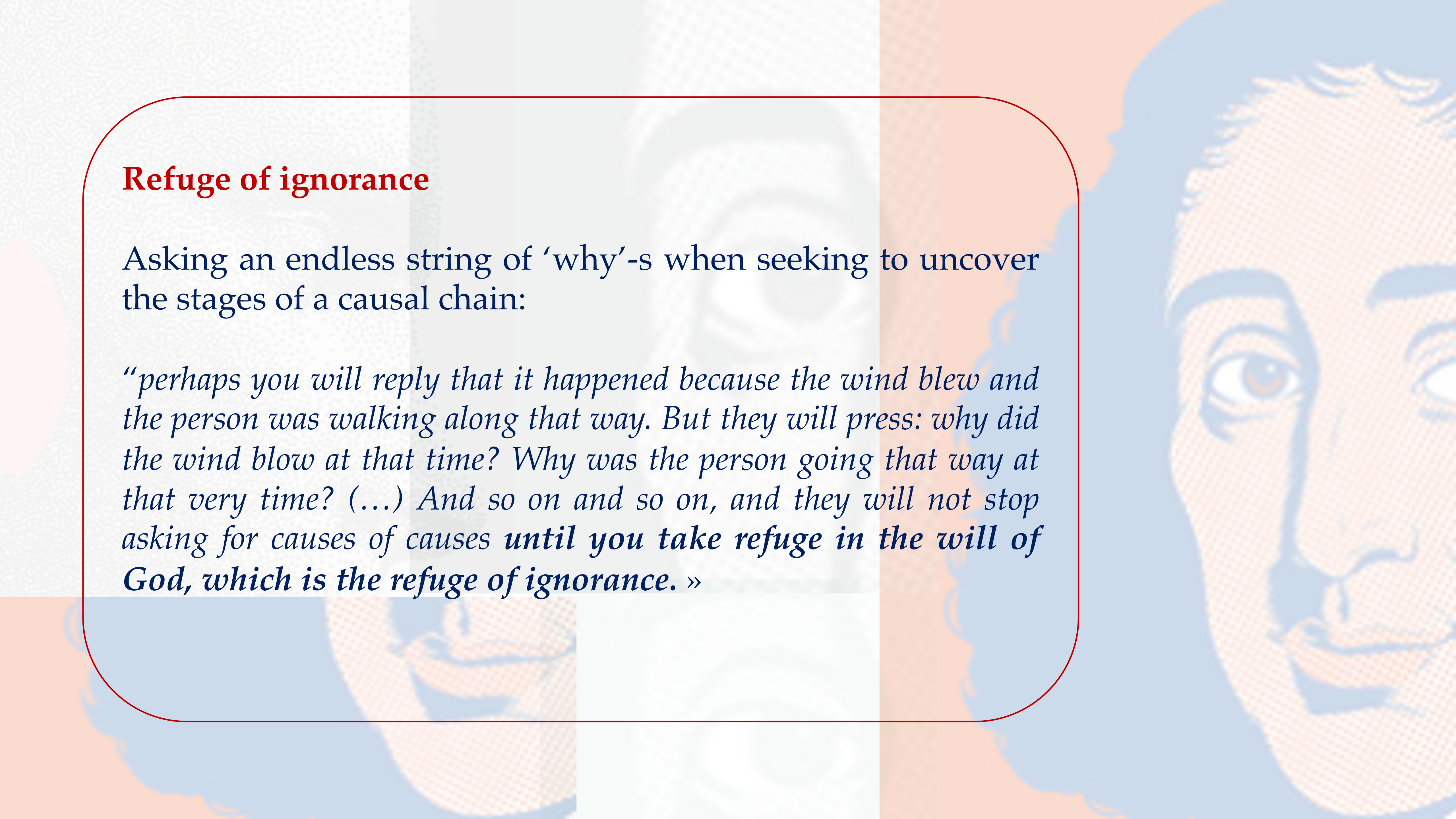
lgrozdanovski@uliege.be

fnrs FREEDOM TO RESEARCH

LIÈGE université

JUST-AI Erasmus+

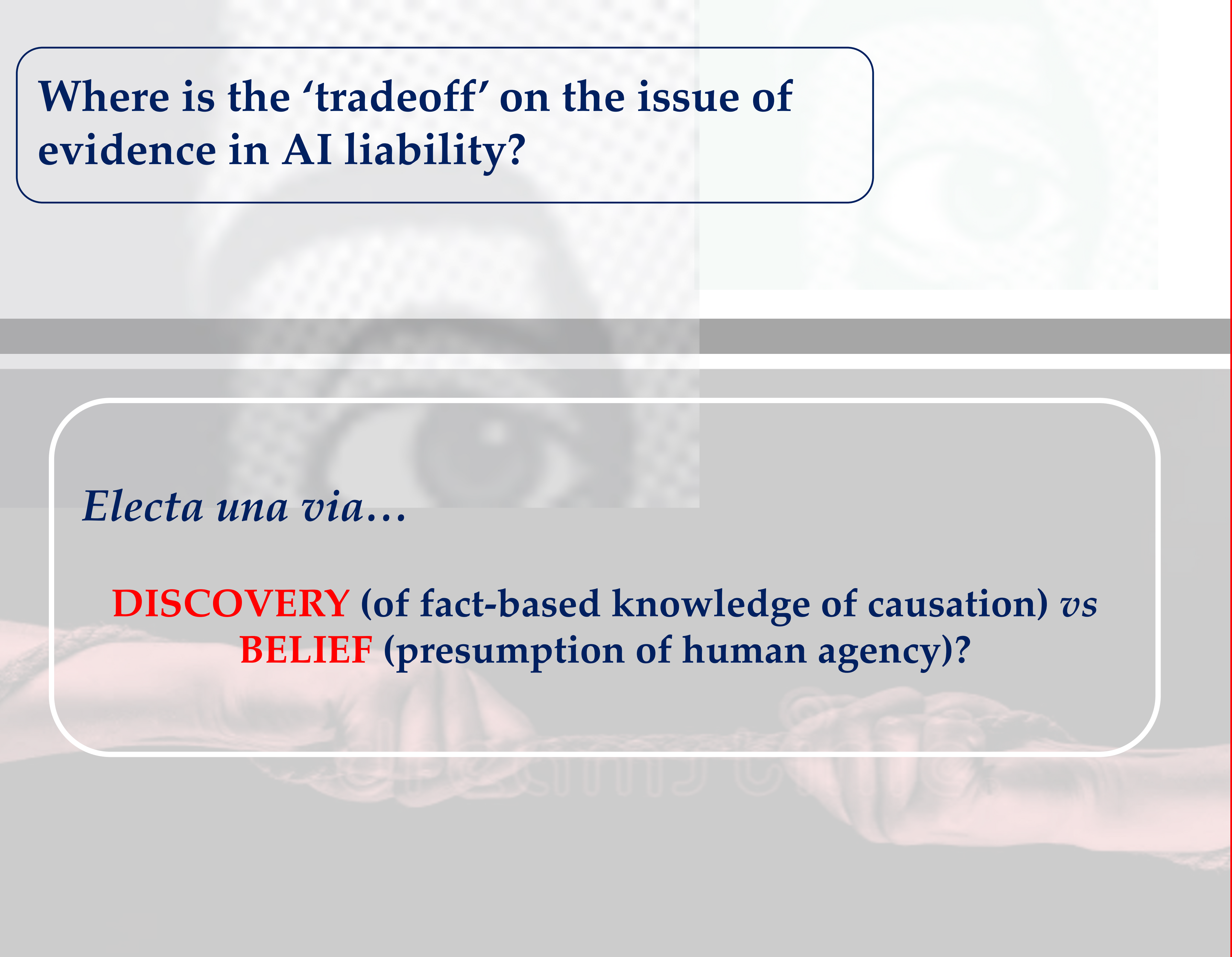
Jean Monnet Center of Excellence



Refuge of ignorance

Asking an endless string of ‘why’-s when seeking to uncover the stages of a causal chain:

“perhaps you will reply that it happened because the wind blew and the person was walking along that way. But they will press: why did the wind blow at that time? Why was the person going that way at that very time? (…) And so on and so on, and they will not stop asking for causes of causes ***until you take refuge in the will of God, which is the refuge of ignorance.*** »



Where is the 'tradeoff' on the issue of evidence in AI liability?

Electa una via…

DISCOVERY (of fact-based knowledge of causation) *vs* BELIEF (presumption of human agency)?

Explainabitliy in the EU's regulatory discourse on AI

HLEG, *Ethics Guidelines* (2019):

HLEG, Ethics Guidelines (2019), at 13:

Explicability is crucial for building and maintaining users' trust in AI systems. This means that **processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions** - to the extent possible - **explainable to those directly and indirectly affected**.

Without such information, a decision cannot be duly contested. An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible (…)

The **degree to which explicability is needed is highly dependent on the context and the severity of the consequences** if that output is erroneous or otherwise inaccurate.

Art. 68(c) AI Act

Any affected person subject to a decision which is taken by the deployer on the **basis of the output from an high-risk AI system which produces legal effects or similarly significantly affects him or her** in a way that they consider to **adversely impact their health, safety, fundamental rights, socio-economic well-being or any other of the rights** deriving from the obligations laid down in this Regulation, shall have the right to **request from the deployer clear and meaningful explanation** pursuant to Article 13(1) on the role of the AI system in the decision-making procedure, the main parameters of the decision taken and the related input data.

Q 1

Is XAI a component of causal explanations (in cases of harm occasioned by AI systems)?

Q 2

If yes (Q 1), does EU law provide de necessary procedural abilities to litigants?

ANALYTICAL FRALEWOERK

PROCEDURAL ABILITIES - 'SPIN OFF' FROM THE CAPABILITIES STRAND (NUSSBAUM (2006); SEN (2009))

Owusu-Bempah (2018):

1) understand the nature of the charge;

2) understand the evidence adduced;

3) understand the trial process and the consequences of being convicted;

4) give instructions to a legal representative;

5) make a decision about whether to plead guilty or not guilty;

6) make a decision about whether to give evidence;

7) make other decisions that might need to be made by the defendant in connection with the trial;

8) follow the proceedings in court on the offence;

9) give evidence;

10) any other ability that appears to the court to be relevant in the particular case.

Accuracy criteria for explanations *tout court*

Explanations are **contrastive**, **selected** and **social**
Michael Ridley, "Explainable Artificial Intelligence (XAI), 41 (2022) 2, *Information technology and libraries*, 1-17, at 4.

Facticity

Believability

Explanations are **context-specific** and **factive**
Andrés Paez," The Pragrmatic Turn in Explainable Artificial Intelligence" (2019) 3 *Minds and machines*, 441, at 454.

Causal explanations

Cause

Necessary and sufficient event for the occurrence of another event

Purpose

Identifying causation (as opposed to correlation)

Wolfgang Pietsch, *On the Epistemology of Data Science* (Springer : 2022), at 110.

Risks

Causal underdetermination / overdetermination

H. L. A. Hart, Tony Honoré, *Causation in the Law* (ed. 1985), at 407.

Tests

Sine qua non* or *but-for (and its variants)

Michael S. Moore, *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics, cit.* (OUP : 2009) , at 83.

Accuracy criteria Explainable AI (XAI)

Barredo Arrieta *et al*. “Expainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” (2020) 58 *Information Fusion*, 82.

Ad hoc

Post hoc

Guidotti *et al*. “Principles of Explainable Artificial Intelligence” in Moamar Sayed-Mouchaweh (ed.), Explainable AI Withiin the Digital Transformation and Cyber Physical Systems: XAI Methods and Applications, Springer (2021), 9.

The go-to evidence: Expertise
Bradford-Hill criteria (for probative scientific evidence)

strength of the causal association / ***consistency*** (stemming from the converging results from different investigations performed in different places) / ***specificity*** (the association should be restricted to a specific cause-effect interrelationship) / *temporal precedence* (the cause must consistently precede the harm) / ***gradient*** (essentially a threshold of gravity) / ***plausibility*** (the cause-effect connection should be plausibly considered as causation) / ***coherence*** (the causal interpretation should not seriously conflict with known facts about the cause-effect interrelationship).

Austin Bradford Hill, "The Environment and Disease: Association or Causation?" *Proceedings of the Royal Society of Medicine*, 58 (1965), 295

The go-to evidence: Expertise?...

Superior Court of New Jersey (Appellate Division), 2 February 2021, *State of New Jersey v. Corey Pickett,* Docket N° A-4207-19T4 (*reverse-engineering of TrueAllele*)

Software program contained approx. 170’000 lines of code written in MATLAB (a programming language designed specifically for visualizing and programming numerical algorithms).

At 207: ‘it would take hours to decipher a few dozen lines of the dense mathematical text comprising the code (amounting to) about ***eight and a half years*** **to review the code in its entirety**.’

How to establish & explain causation in AI liability – knowledge or belief?

Design of the AI Liability framework in the EU

PROCEDURAL ABILITIES IN THE AILD / R-PLD

The Right

The right to request disclosre of evidence
Art. 3 AILD, Art. 8 R-PLD

L. Grozdanovski, 'In search for effetiveness and fairness in proving algorithmic discrimination in EU law' CMLRev. 58-1 (2021)

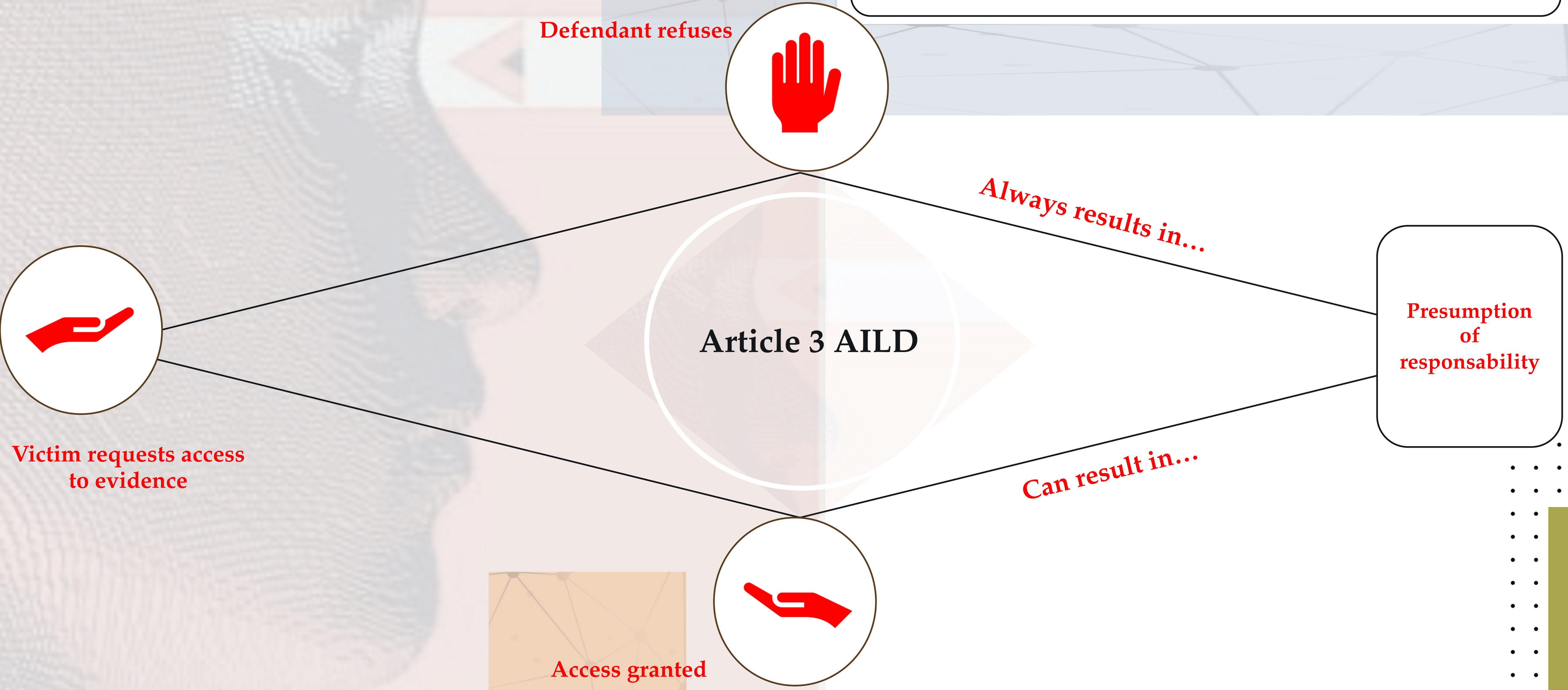
A 'web' of presumptions

Art. 3 AILD – presumption of fault/responsability

Art. 8 PLD – presumption of defectiveness

The Effects of the right

How does it work?



Article 3 AILD

Victim requests access to evidence

Defendant refuses

Always results in…

Access granted

Can result in…

Presumption of responsability

How does it work?

Presumption of defectiveness

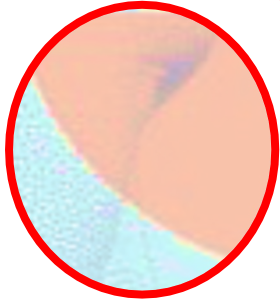
1°) The defendant had not respond favorably to the request to disclose evidence

2°) The 'product' does not comply with mandatory technical standards

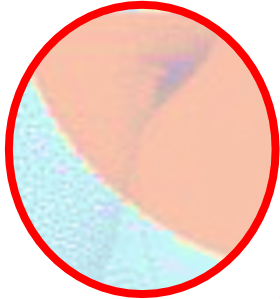
3°) Harm occurred due to a manifest malfunction, given the product's 'normal' use / ordinary circumstances

The cricitism: the evidentiary hermetism of the AILD/R-PLD

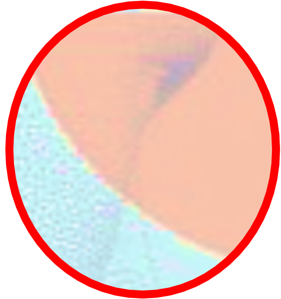
Evidentiary debates are limited to *ad hoc* explainabiloity



Incoherences in the application of the presumptions of fault/defectiveness. What of defendants?



What if harm results from lawful conduct?



ARE *POST HOC* EXPLANATIONS NECESSARY FOR AI CAUSAL EXPLANATIONS?

LESSONS FROM THE EMERGING CASELAW IN AI LIABILITY

Superior Court of New Jersey (Appellate Division), 2 février 2021, *State of New Jersey v. Corey Pickett*, Docket N° A-4207-19T4

Supreme Court of Wisconsin, 13 juillet 2016 (decided), *State of Wisconsin v. Eric L. Loomis*, 881 N.W. 2d 749 (2016) 2016 WI 68

Ewert vs. Canada, 2018 SCC 30, File n° 37233, 13 juin 2018

Victims always request *post hoc* explanations

Courts (almost always) request independent expertise

Victims seek to understand the reasons for (automatic) human reliance on AI output

Moral of the story

From procedural abilities…

NO ABILITIES in view of giving/receiving *post hoc* explanations

…through the design of systems of evidence

Systems do not seem (overly) permissible to evidence flagged as necessary (e.g. expertise)

…to a theory of 'AI procedural justice'

Procedural justice requires that causal (AI) explanations integrate 'full' XAI (*ad hoc* & *post hoc*)

Closing the circle…

Are we in a ‘refuge of ignorance’ the EU legislature built (through the AILD/R-PLD)?