Winter Term 21/22

Artificial Intelligence, Ethics & Engineering

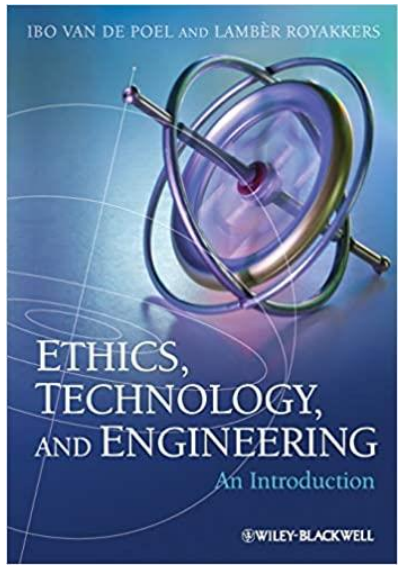
Lecture-3: Responsibilities & Codes of Conduct

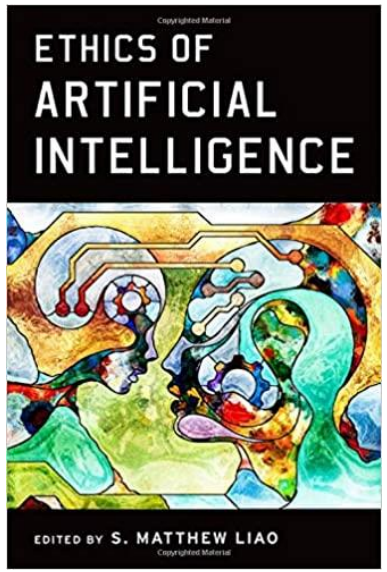
Prof. Dr. Holger Giese (holger.giese@hpi.uni-potsdam.de)

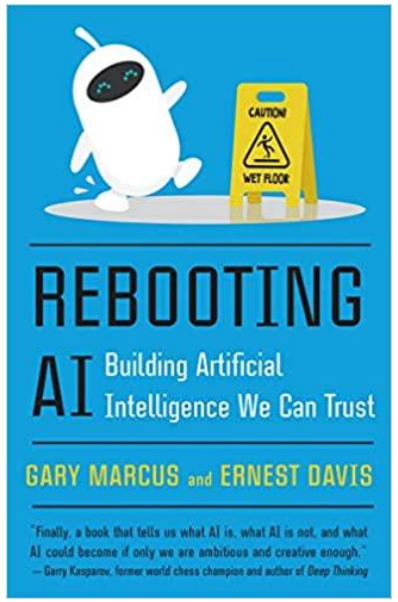
Christian Medeiros Adriano (christian.adriano@hpi.de) - **"Chris"**

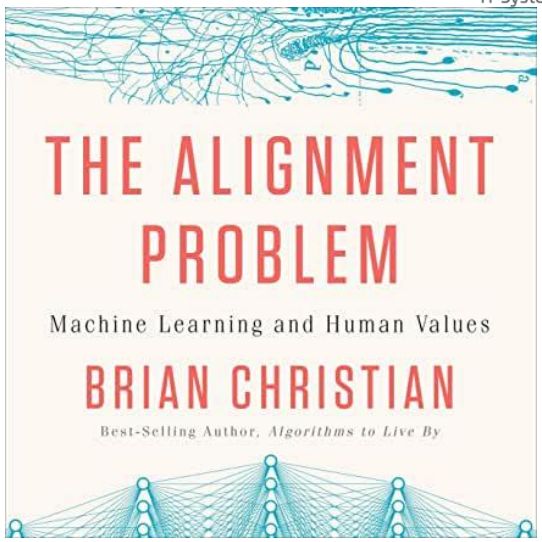
Christian Zöllner (Christian.zoellner@hpi.de)

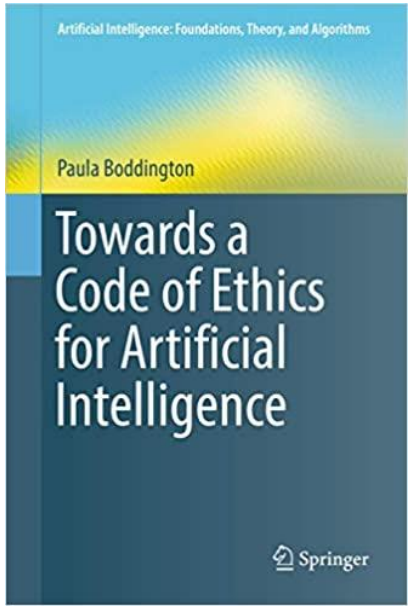
Sources (Books)

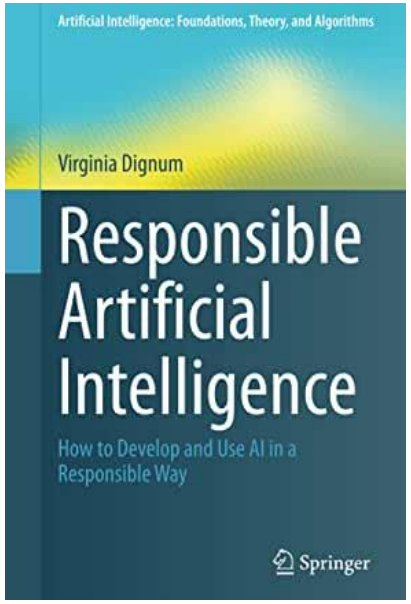


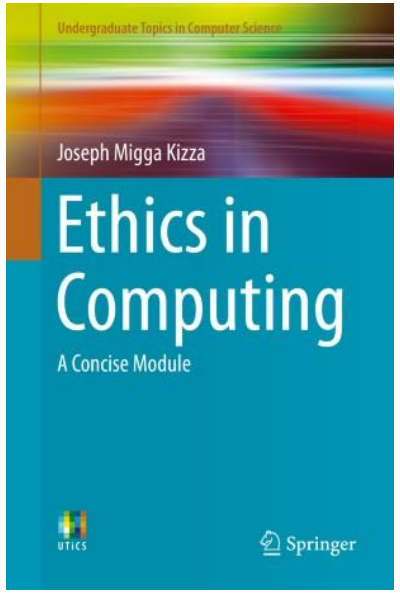


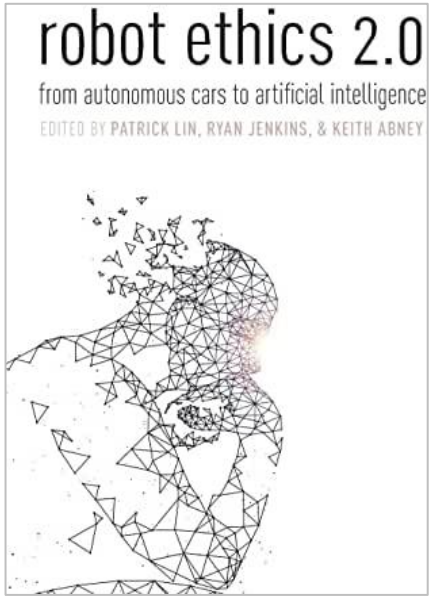












Responsibility

- In case of a **disaster** or **system failure** the question who is responsible naturally arises!
- **Active responsibility**: before the event
- **Passive responsibility**: after the event
- Different roles result in different responsibilities (engineer vs. employee)
- Responsibilities can result from contracts, code of conducts, or moral norms and moral duties

[dePoel&Royakkers2011]

[dePoel&Royakkers2011] Van de Poel, I., & Royakkers, L. (2011). *Ethics, technology, and engineering: An introduction*. John Wiley & Sons.

Kind of Responsibilities

- **Role responsibility**: The responsibility that is based on the role one has or plays in a certain situation.
- **Moral responsibility**: Responsibility that is based on moral obligations, moral norms or moral duties.
- **Professional responsibility**: The responsibility that is based on one's role as professional in as far it stays within the limits of what is morally allowed.

[dePoel&Royakkers2011]

Distribution of Responsibility

- **Collective Responsibility**: The responsibility of a collective of people.
- **Problem of many hands**: The occurrence of the situation in which the collective can reasonably be held morally responsible for an outcome, while none of the individuals can be held responsible for that outcome.
- **Distribution of Responsibility**: The ascription or apportioning of (individual) responsibilities to various actors.
  - **Moral Fairness Requirement**: The requirement that a distribution of responsibility should be fair (see Blameworthiness).
  - **Effectiveness Requirement**: The moral requirement that states that responsibility should be so distributed that the best consequences, that is, is effective in preventing harm (and in achieving positive consequences).

[dePoel&Royakkers2011]

Passive Responsibility

Backward-looking responsibility, relevant after something occurred; specific forms are:

- **Accountability**: Backward-looking responsibility in the sense of being held to account for or justify one's action towards others.
- **Blameworthiness**: Backward-looking responsibility in the sense of being a proper target for blame for one's action or its consequences. The following conditions need to apply:
  - wrong-doing
  - causal contribution
  - foreseeability, and
  - Freedom of action.
- **Liability / Legal Responsibility**: Backward-looking responsibility according to the law.

[dePoel&Royakkers2011]

Moral Responsibility vs. Liability

Moral Responsibility	Legal Liability
Blameworthiness (wrong-doing, causality, foreseeability, freedom)	Based on conditions formulated in **laws**
Informally established	Formally established in court
Not necessarily connected to punishment or compensation	Implies obligation to pay a fine or to repay damages
Backward-looking and **forward-looking**	Only backward-looking

[dePoel&Royakkers2011]

Regulation

Other tools to deal with social consequences of technology:

- **Regulation**: A legal tool that can forbid the development, production, or use of certain technological products, but more often it formulates a set of the boundary conditions for the design, production, and use of technologies.

[dePoel&Royakkers2011]

Negligence vs. Strict Liability

- **Negligence**: Not living by certain duties. Negligence is often a main condition for legal liability. In order to show negligence for the law, usually proof must be given of a duty owed, a breach of that duty, an injury or damage, a causal connection between the breach and the injury or damage.

- **Duty of Care**: The legal obligation to adhere to a reasonable standard of care when performing any acts that could foreseeably harm others.

- **Strict Liability**: A form of liability that does not require the defendant to be negligent.

[dePoel&Royakkers2011]

Product Liability and its Limits ...

- **Product Liability**: Liability of manufactures for defects in a product, without the need to proof that those manufactures acted negligently.

BUT: exception for development defects (e.g., EU directives):

- **Development Risk**: In the context of product liability : Risk that could not have been foreseen given the **state of scientific and technical knowledge** at the time the product was put into circulation.

- **Corporate Liability**: Liability of a company (corporation) when it is treated as a legal person.

BUT: liability of a corporation may be limited:

- **Limited Liability**: The principle that the liability of shareholders for the cooperation's debts and obligations is limited to the value of their shares.

[dePoel&Royakkers2011]

Active Responsibility

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Responsibility before something has happened referring to a duty or task to care for certain state-of-affairs or person.

Features (according to [Bovens1998]):

- Adequate perception of threatened violations of norms;
- Consideration of the consequences;
- Autonomy, i.e., the ability to make one's own independent moral decisions;
- Displaying conduct that is based on a verifiable and consistent code; and
- Taking role obligations seriously.

[dePoel&Royakkers2011]

Active Responsibility & Engineering Ideals

- **Ideals** or strivings which are particularly motivating and inspiring for the person having them, and which aim at achieving an optimum or maximum.
- **Professional ideals** are closely aligned to a profession and can only be aspired to by carrying out the profession.
  - **Technological enthusiasm**: The ideal of wanting to develop new technological possibilities and taking up technological challenges. Not morally improper, but leads to easily overlook moral issues ...

    (Wernher von Braun (1912-1977)
  - **Effectiveness** is the extent to which an established goal is achieved, and **efficiency** is the ratio between the goal achieved and the effort required. But the goal may be not morally justified ...

    (Frederick W. Taylor (1856-1915), Adolf Eichmann (1906-1962+))
  - **Human welfare**: It is for sure laudable, but is a moral obligation for all engineers? Shows that engineering is not morally neutral!

[dePoel&Royakkers2011]

Active Responsibility: Engineers vs. Managers

■**Separatism**: The notion that scientists and engineers should apply the technical inputs, but appropriate management and political organs should make the value decisions.

Tripartite model:

- **"Hired gun"**: Someone who is willing to carry out any task or assignment from his employer without any morel scruples.

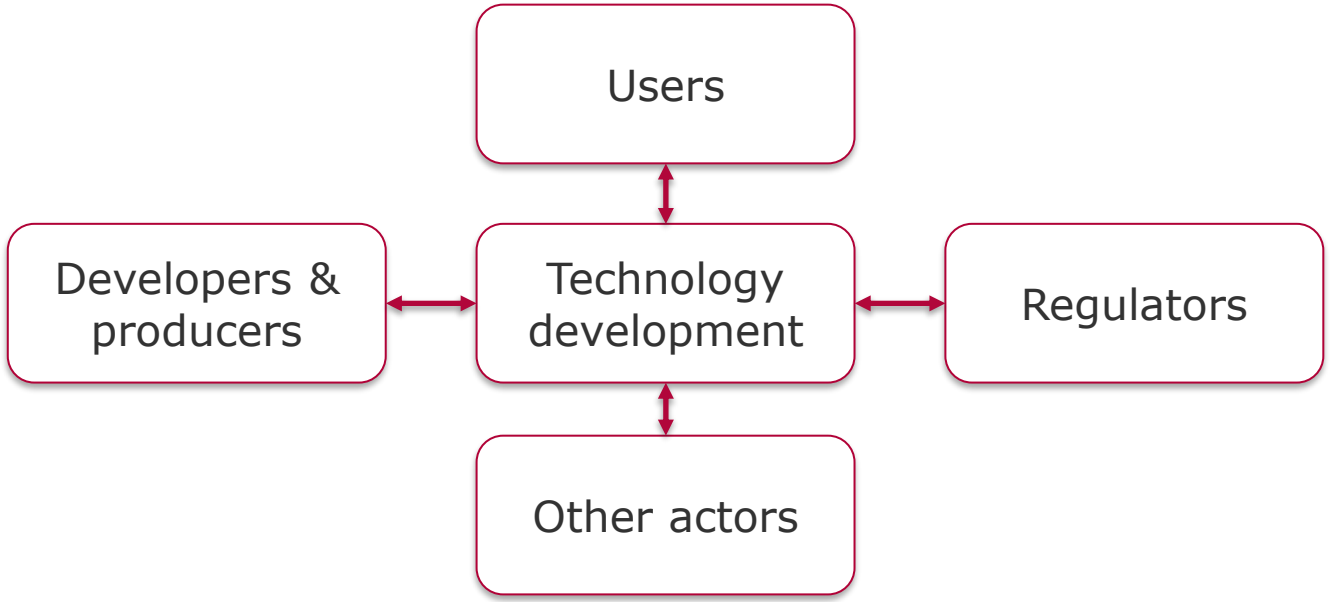
■**Technocracy**: Government by experts.
Engineers should take over the management, but

- **Technocratic Fallacy:** When it comes to the underlying goals or acceptable risk they are not more knowledgeable than others.
- **Paternalism**: The making of (moral) decisions for others on the assumption that one knows better what is good for them than those others themselves.

[dePoel&Royakkers2011]

Active Responsibility: The Social Context

[dePoel&Royakkers2011]



- **Regulators**: Organizations who formulate rules or regulations that engineering products have to meet such as ruling concerning health and safety, but also rulings linked to relations between competitors.
- **Interests**: Action that actors strive for because they are beneficial or advantageous for them.
- **Stakeholders**: Actors that have an interest ("a stake") in the development of a technology.

Active Responsibility: Technology Assessment

- **Technology Assessment (TA)**: Systematic method for exploring future technology developments and assessing their potential societal consequences.

- **Collingridge Dilemma**: The double-blind problem to control the direction of technological development. On the one hand, it is often not possible to predict the consequences of new technology early on. On the other hand, once the (negative) consequences materialize, it often has become difficult to change the direction.

- **Constructive Technology Assessment (CTA)**: TA approach in which TA-like efforts are carried out parallel to the process of technological development and are fed back to the development and design process.

[dePoel&Royakkers2011]

Active Responsibility: Whistle-Blowing

Whistle-Blowing: The disclosure of certain abuses in a company by an employee in which he or she is employed, without the consent of his/her superiors, and in order to remedy these abuses and/or to warn the public about these abuses. Guidelines:

- Reason must be to prevent serious and considerable harm to the public
- The whistle-blower has identified the threat of harm, reported it to its superiors making clear the treat, and concluded that the superior will nothing effective.
- The whistle-blower has exhausted other internal procedures within the organization (as the danger to others and her own safety make reasonable)
- The whistle-blower has evidence that convince a reasonable, impartial observer that her view of the threat is correct.
- The whistle-blower has good reasons to believe that revealing the threat will (probably) prevent the harm at reasonable cost.

[dePoel&Royakkers2011]

Codes of Conduct

- **Codes of Conduct**: A code in which organizations lay down guidelines for responsible behavior of their members. Types:
- **Professional Code**: Code of conduct that is formulated by a professional association (e.g., IEEE, ACM, ...).
  - IEEE-CS/ACM Joint Task Force on Software Engineering Ethics and Professional Practices: Software Engineering Code of Ethics
    https://www.computer.org/web/education/code-of-ethics
  - IEEE https://www.ieee.org/about/corporate/governance/p7-8.html
  - ACM https://www.acm.org/about-acm/acm-code-of-ethics-and-professional-conduct
- **Corporate Code**: Code of conduct that is formulated by a company.
  - Tesla's Code of Business Conduct and Ethics
    http://ir.tesla.com/corporate-governance-document.cfm?documentid=7159
- **Global Code**: Code of conduct that is believed to apply worldwide.
  - The Ten Principles of the UN Global Compact
    https://www.unglobalcompact.org/what-is-gc/mission/principles

[dePoel&Royakkers2011]

Education and Licensing

- **Formal Education**: include material in the curriculum [Kizza2013]
  - Ensuring that individuals posses the necessary technical skills
  - Enforcing that individuals understand the codes of conduct
- **Licensing**: grants individuals formal and legal permission to practice their profession
  - Can be bound to a certain formal education
  - Can require to demonstrate skill in dedicated tests
  - Ensuring that professionals posses the necessary technical skills
  - Enforcing that professionals understand the codes of conduct

HOWEVER: The software engineering community has neither established that **formal education** nor **licensing** are prerequisites to act as a software engineering professional in related projects.
(ACM abandon related efforts, IEEE continues them)

[Kizza2013] Kizza, J. M. (2016). *Ethics in computing*. Springer.

EU regulation [1][2] – Uses of AI

The proposal sets a nuanced regulatory structure that

- bans some uses of AI
  - Paragraph 23 - The use of AI systems for 'real-time' remote biometric identification of natural persons in publicly accessible spaces for the purpose of law enforcement necessarily involves the processing of biometric data.
- heavily regulates high-risk uses
  - rules on data and data governance; documentation and record-keeping; transparency and provision of information to users; human oversight; and robustness, accuracy and security.
- lightly regulates less risky AI systems

sources:
[1] MacCarthy, M., & Propp, K., 2021, Machines learn that Brussels writes the rules: The EU's new AI regulation, https://www.brookings.edu/blog/techtank/2021/05/04/machines-learn-that-brussels-writes-the-rules-the-eus-new-ai-regulation/
[2] European Commission, 2021, Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS, https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

EU regulation [1][2] – Robustness for High-Risk AI

- has “proportionate” postmarket monitoring system to collect data on the system’s operation to ensure its “continuous compliance”
- be “sufficiently transparent to enable users to understand and control how the high-risk AI system produces its output.”
- discloses its “the level of accuracy, robustness and security,”
- “meet a high level of accuracy that is appropriate for their intended purpose” and to continue to perform at that level of accuracy in use.
- be resilient against “errors, faults or inconsistencies” and also against “attempts to alter their use or performance by malicious third parties intending to exploit system vulnerabilities.”

sources:
[1] MacCarthy, M., & Propp, K., 2021, Machines learn that Brussels writes the rules: The EU’s new AI regulation, https://www.brookings.edu/blog/techtank/2021/05/04/machines-learn-that-brussels-writes-the-rules-the-eus-new-ai-regulation/
[2] European Commission, 2021, Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS, https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

EU regulation [1][2] – Providers

- The providers of AI systems used as safety components of consumer products, who are already subject to third-party ex-ante conformity assessment under current product safety law, now must also demonstrate compliance with the AI Act [2].

- conduct conformity assessments demonstrating that the high-risk system complies with these rules.

sources:
[1] MacCarthy, M., & Propp, K., 2021, Machines learn that Brussels writes the rules: The EU's new AI regulation, https://www.brookings.edu/blog/techtank/2021/05/04/machines-learn-that-brussels-writes-the-rules-the-eus-new-ai-regulation/
[2] European Commission, 2021, Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS, https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

EU regulation [1][2] – Gaps?

- People should be informed when they "interact with" an AI system or when their emotions or gender, race, ethnicity or sexual orientation are "recognized" by an AI system.
- They must be told when "deepfake" systems artificially create or manipulate material.

However,
- no requirements to inform people when thy are subjected to algorithmic assessments.
- no goals for algorithmic fairness (although it was discussed in meetings)
- required conformity assessments cover only internal processes, i.e., not documents that could be reviewed by the public or a regulator.

sources:
[1] MacCarthy, M., & Propp, K., 2021, Machines learn that Brussels writes the rules: The EU's new AI regulation, https://www.brookings.edu/blog/techtank/2021/05/04/machines-learn-that-brussels-writes-the-rules-the-eus-new-ai-regulation/
[2] European Commission, 2021, Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS, https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

Informal Definition/Aspirations [R²AW 2021]

Topics for Robust and Reliable Autonomy in the Wild

- Definitions of safety, robustness, reliability, and resilience
- Evaluation metrics for robustness and reliability, under model imprecision
- Decision-making representations, models, and algorithms for the open world
- Techniques to achieve resilient decision-making, under unmodelled disturbances
- Techniques to recognize and avoid negative side effects of AI systems
- Techniques for ethical, interpretable, fair, and trustworthy decision-making
- Case studies of robustness and reliability in deployed autonomous systems
- Learning to improve robustness and reliability from human feedback

R²AW, 2021, IJCAI 2021 Workshop on Robust and Reliable Autonomy in the Wild, http://rbr.cs.umass.edu/r2aw/

Suggested Task for the Week

Read the IEEE code of Ethics [1], IEEE code of Conduct [2], and the EU Rules on AI (AI Act) [2-3]

Are there specific orientations in these codes that contribute to mitigate the dilemma that you described in your previous tasks?

- If positive, please explain (2 sentences) how.
- If negative, what would you recommend to be added or stated more precise/specific in these codes in order to contribute to a particular mitigation action?

[1] IEEE Code of Ethics, 2020, https://www.ieee.org/content/dam/ieee-org/ieee/web/org/about/corporate/ieee-code-of-ethics.pdf

[2] IEEE Code of Conduct, 2014, https://www.ieee.org/content/dam/ieee-org/ieee/web/org/about/ieee_code_of_conduct.pdf

[3] MacCarthy, M., & Propp, K., 2021, Machines learn that Brussels writes the rules: The EU's new AI regulation, https://www.brookings.edu/blog/techtank/2021/05/04/machines-learn-that-brussels-writes-the-rules-the-eus-new-ai-regulation/

[4] European Commission, 2021, Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS, https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

End