Ethics of Big Data

Yves Moreau

KU LEUVEN

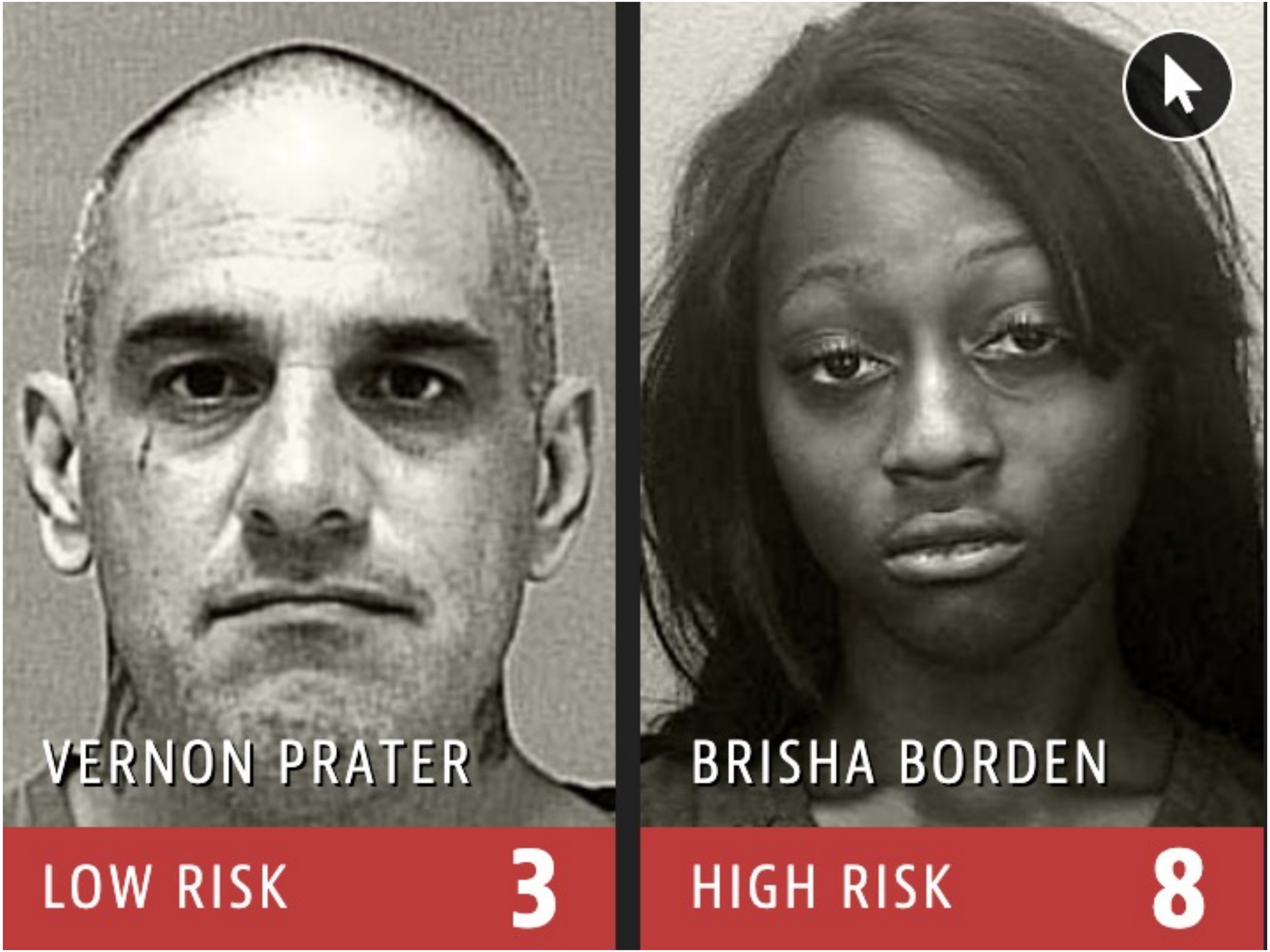
Weapons of math destruction

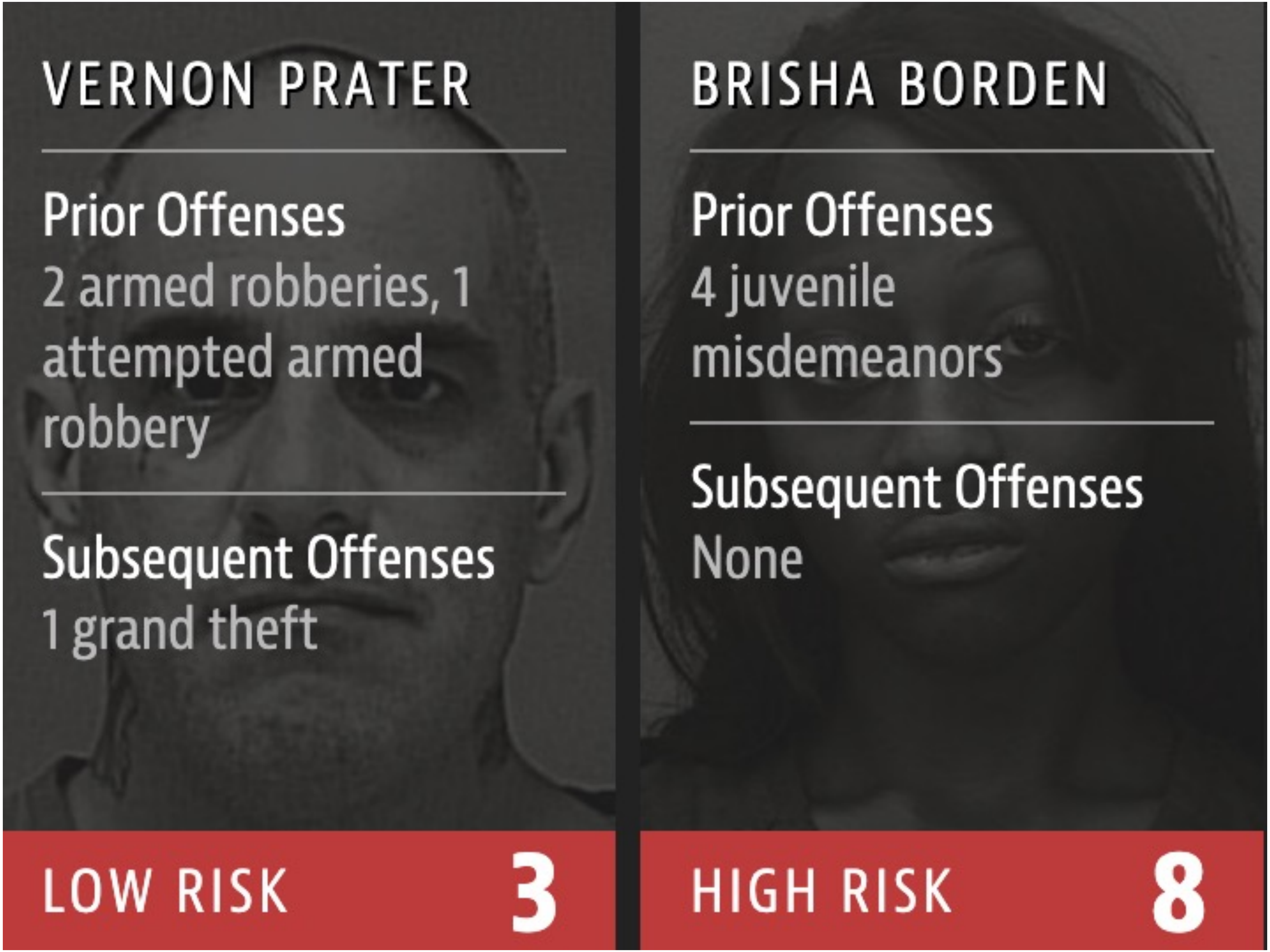
Weapons of math destruction

Cathy O’Neil. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group New York, NY, USA, 2016.

- https://www.youtube.com/watch?v=TQHs8SA1qpk (60’)

1. Model is widespread and high stakes
2. Model is secret
3. Data is biased
4. Covariates are proxies for unethical biases
5. Measure of success is questionable
6. Model creates vicious circles


VERNON PRATER
LOW RISK
3
BRISHA BORDEN
HIGH RISK
8



VERNON PRATER

Prior Offenses

2 armed robberies, 1 attempted armed robbery

Subsequent Offenses

1 grand theft

LOW RISK 3

BRISHA BORDEN

Prior Offenses

4 juvenile misdemeanors

Subsequent Offenses

None

HIGH RISK 8

Social enforcement

- Pseudo-objectivity
  - “It’s math”
- Mathematical models as argument of authority
  - “It’s math and you wouldn’t understand it”
- Bureaucratic displacement of responsibility
- Bureaucratic rules as social gatekeeping

Biases

- Data can be biased
  - Current data will reflect current biases so that model will reproduce those biases
- Covariates can be correlated with variables that are morally or legally unacceptable to use
  - *e.g.*, ZIP code as a proxy for wealth and race
- Objective function reflects a value system
  - In practice, big gap between general goal (“make people safe”) and specific objective function (“predict misdemeanor arrest”)

Data Science Ethics Checklist

- http://deon.drivendata.org/

A. Data Collection

- A.1 Informed consent: If there are human subjects, have those subjects have given informed consent, where users clearly understand what they are consenting to and there was a mechanism in place for gathering consent?
- A.2 Collection bias: Have we considered sources of bias that could be introduced during data collection and survey design and taken steps to mitigate those?
- A.3 Limit PII exposure: Have we considered ways to to minimize exposure of personally identifiable information (PII) for example through anonymization or not collecting information that isn't relevant for analysis?

B. Data Storage

- B.1 Data security: Do we have a plan to protect and secure data (e.g., encryption at rest and in transit, access controls on internal users and third parties, access logs, and up-to-date software)?
- B.2 Right to be forgotten: Do we have a mechanism through which an individual can request their personal information be removed?
- B.3 Data retention plan: Is there a schedule or plan to delete the data after it is no longer needed?

C. Analysis

- C.1 Missing perspectives: Have we sought to address blindspots in the analysis through engagement with relevant stakeholders (e.g., checking assumptions and discussing implications with affected communities and subject matter experts)?
- C.2 Dataset bias: Have we examined the data for possible sources of bias and taken steps to mitigate or address these biases (e.g., stereotype perpetuation, confirmation bias, imbalanced classes, or omitted confounding variables)?
- C.3 Honest representation: Are our visualizations, summary statistics, and reports designed to honestly represent the underlying data?

C. Analysis

- C.4 Privacy in analysis: Have we ensured that data with PII are not used or displayed unless necessary for the analysis?
- C.5 Auditability: Is the process of generating the analysis well documented and reproducible if we discover issues in the future?

D. Modeling

- D.1 Proxy discrimination: Have we ensured that the model does not rely on variables or proxies for variables that are unfairly discriminatory?
- D.2 Fairness across groups: Have we tested model results for fairness with respect to different affected groups (e.g., tested for disparate error rates)?
- D.3 Metric selection: Have we considered the effects of optimizing for our defined metrics and considered additional metrics?

D. Modeling

- D.4 Explainability: Can we explain in understandable terms a decision the model made in cases where a justification is needed?
- D.5 Communicate bias: Have we communicated the shortcomings, limitations, and biases of the model to relevant stakeholders in ways that can be generally understood?

E. Deployment

- E.1 Redress: Have we discussed with our organization a plan for response if users are harmed by the results (e.g., how does the data science team evaluate these cases and update analysis and models to prevent future harm)?
- E.2 Roll back: Is there a way to turn off or roll back the model in production if necessary?
- E.3 Concept drift: Do we test and monitor for concept drift to ensure the model remains fair over time?
- E.4 Unintended use: Have we taken steps to identify and prevent unintended uses and abuse of the model and do we have a plan to monitor these once the model is deployed?

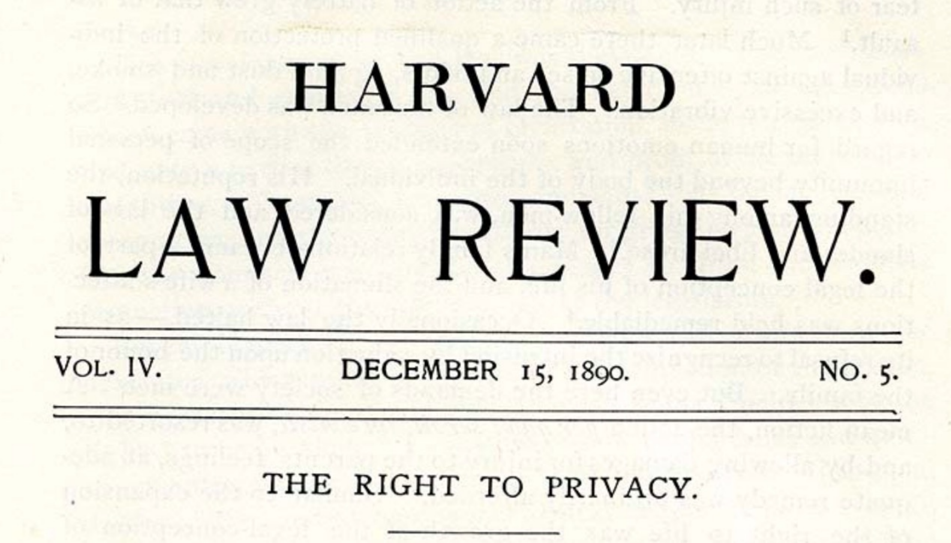
General Data Protection Regulation

Why?

- “If you have nothing to hide, you have nothing to fear”
  - Self-censorship
  - Loss of autonomy
- Is privacy dead?

The right “to be let alone”

- *The right to privacy.* Warren and Brandeis (1890)



HARVARD

LAW REVIEW.

VOL. IV. DECEMBER 15, 1890. NO. 5.

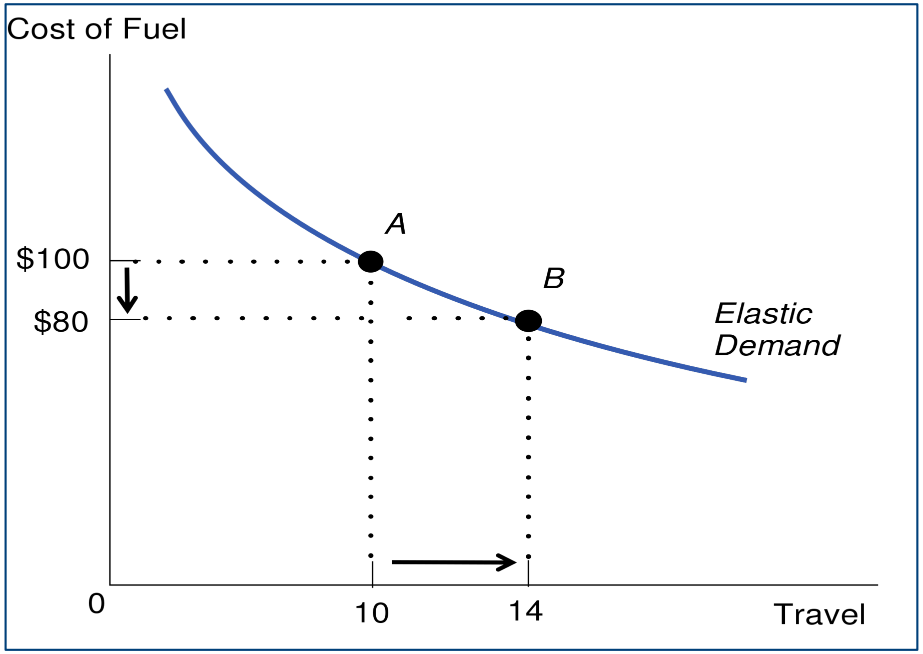
THE RIGHT TO PRIVACY.

- Privacy is key to autonomy
- Privacy is an individual human right and a common good
- NOYB: None of your business

Economics of surveillance

- Jevons' paradox

- Video surveillance
  - Low marginal cost
- Facial recognition
  - Near zero marginal cost



- Without constraint, near infinite demand for surveillance

General Data Protection Regulation

- Previous legislation: Data Privacy Directive (1995)
  - Directive: specific implementation per country
- General Data Protection Regulation (2018)
  - Regulation: harmonized across the EU
  - Some aspects interpreted at national level
  - Increased protections/rights
  - Much stronger sanctions that DPD
- Applies to organizations that hold data on EU citizens and residents (also non-EU organizations)

Data controllers and processors

- GDPR applies to Controllers (say how and why data is processed) and Processors (process data on behalf of controllers)

- A data controller is the individual or the legal person who controls and is responsible for the keeping and use of personal information on computer or in structured manual files.
- A data processor is anyone who processes personal data **on behalf of** the data controller (excluding the data controller's own employees). For example, storage of the data on a third party's servers, or appointing a data analytics provider.

Personal data

- GDPR applies to personal data only
  - Does not apply to anonymous data, although other legislation might be relevant
- Personal data
  - Not only data linked to name and address
  - Includes online identifiers (e.g., cookies or IP address)
  - Any information that can be linked back to a unique person using plausible means
    - If someone else can de-anonymize the data with reasonable effort, it is personal data
  - Patient genomic data is personal in and of itself

Personal data

- Sensitive Personal Data
  - Generally, sensitive information about an individual
    - Race/ethnicity, religion, politics, trade unions, sex life
    - Health, genetics, biometrics
- Special rules for processing children data

- ***The GDPR does not forbid processing personal data!***
  - How to process it safely

GDPR principles

- Lawful processing
- Data collected for a specific, legitimate purpose
- Adequate, relevant and limited to that purpose
- Accurate and kept up to date
- Kept for no longer than needed
- Kept secure

- Much enhanced principle of ACCOUNTABILITY

Accountability

- Critical new principle
- Organizations must DEMONSTRATE compliance
  - Documenting processing activities
  - Appoint a Data Protection Officer?
  - Data protection impact assessments
  - Data protection “by design and by default”
  - Maintain records of processing activities

- Must actively demonstrate compliance

Basis for processing

- Have to demonstrate a legal basis for processing
- This can include
  - Consent
  - Legitimate basis for processing (including performance of a contract)
  - Public interest

- ***Importantly, consent is not the only acceptable basis for processing***

Individual rights

- Enhanced existing rights
  - Right to be informed
  - Right of access
  - Right of rectification
  - Right to object
  - Rights regarding automated processing
- New rights
  - Right to restriction
  - Right to erasure
  - Right to data portability

Consent

- Important – consent is not the only acceptable legal basis for processing personal data
  - If consent is basis for processing sensitive personal data, consent MUST be explicit
- Consent requires "clear, affirmative action" (*i.e.*, not a pre-ticked box)
- It must be freely given, informed, specific, and verifiable
- It can be withdrawn at any time

Breach notification and enforcement

- Breaches generally expected to be report within 72 hours (but also ‘without undue delay’)
- Extends mandatory breach reporting beyond ISPs and telcos to all controllers/processors
- Report to data controllers, regulators and – in some cases – affected data subjects

- FINES – up to €20m or 4% of global turnover for major breaches
- Up to €10m or 2% of global turnover for minor breaches

Profiling

- Any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyze or predict aspects concerning that natural person’s performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements (GDPR Art. 4)

Profiling

- Processing shall be lawful only if and to the extent that at least one of the following applies
  - Consent for one or more specific purposes
  - Performance of a contract
  - Compliance with a legal obligation
  - Necessary to protect the vital interests of the data subject or of another natural person
  - Public interest or official authority
  - Legitimate interests of controller, except when overridden by interests of data subject

Automated decision-making

- Individual has right not to be subject to a decision based solely on automated processing
  - Profiling is not in and of itself an automated decision!
  - There must be a decision
  - There must be automated processing (which may include profiling)
  - Decision must be based solely on automated processing
  - Decision must produce “legal effects” or otherwise “significantly affect” the individual

Automated decision-making

- Automated decision making IS permitted if
  - Authorized by Union or Member State law
  - Necessary for the contract between the data subject and data controller
  - Data subject has provided explicit consent.
- BUT
  - Right to express their view
  - ***Right to obtain explanation of decision reached***
  - Right to object / challenge the decision
  - Sensitive data / children

Automated decision-making

- Ensure data is processed fairly and transparently
  - Use appropriate mathematical or statistical procedures
  - Implement technical and organizational measures to avoid and correct errors
  - Minimise bias or discrimination
  - Provide meaningful clear information (1) about existence of automated decision making, including profiling and (2) logic involved and significance and envisaged consequences of profiling.

Automated decision-making

- Comply with principles of accuracy, storage limitation and privacy by design
  - Data must be kept accurate and up-to-date – garbage in, garbage out?
  - Ensure data is not kept for longer than necessary
  - Incorporate processes by default and by design
- Honor the “right to object” exercised by any data subject (whether or not automated)
- Carry out Data Protection Impact Assessment (DPIA) for high risk processing
- Appoint Data Protection Officer (DPO) if required

Research exemption

- Exemptions easing secondary use of data for research
- Major lobbying by scientific organizations to push back against restrictive proposal
  - Must avoid abuses and contribute to public good
- Article 5 principles relating to personal data processing
  - (b) collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes; **further processing of personal data for archiving purposes in the public interest or *scientific*, statistical or historical purposes shall in accordance with Article 89 not be considered incompatible with the initial purposes.**

Research exemption

- (e) kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed (…); **personal data may be stored for longer periods insofar as the data will be processed for archiving purposes in the public interest or scientific, statistical, or historical purposes in accordance with Article 89 (1)**

Research exemption

- Article 9 (j) Processing of special categories of personal data is prohibited
  - Unless the data subject has given explicit consent. (freely given, informed, specific and unambiguous)
  - The processing relates to personal data which are manifestly made public by the data subject
  - Processing is necessary for archiving purposes in the public interest, or scientific and historical research purposes according with Article 89 (1)
- Member states may maintain or introduce further conditions, with regard to the processing of genetic data, biometric data or health data

Research exemption

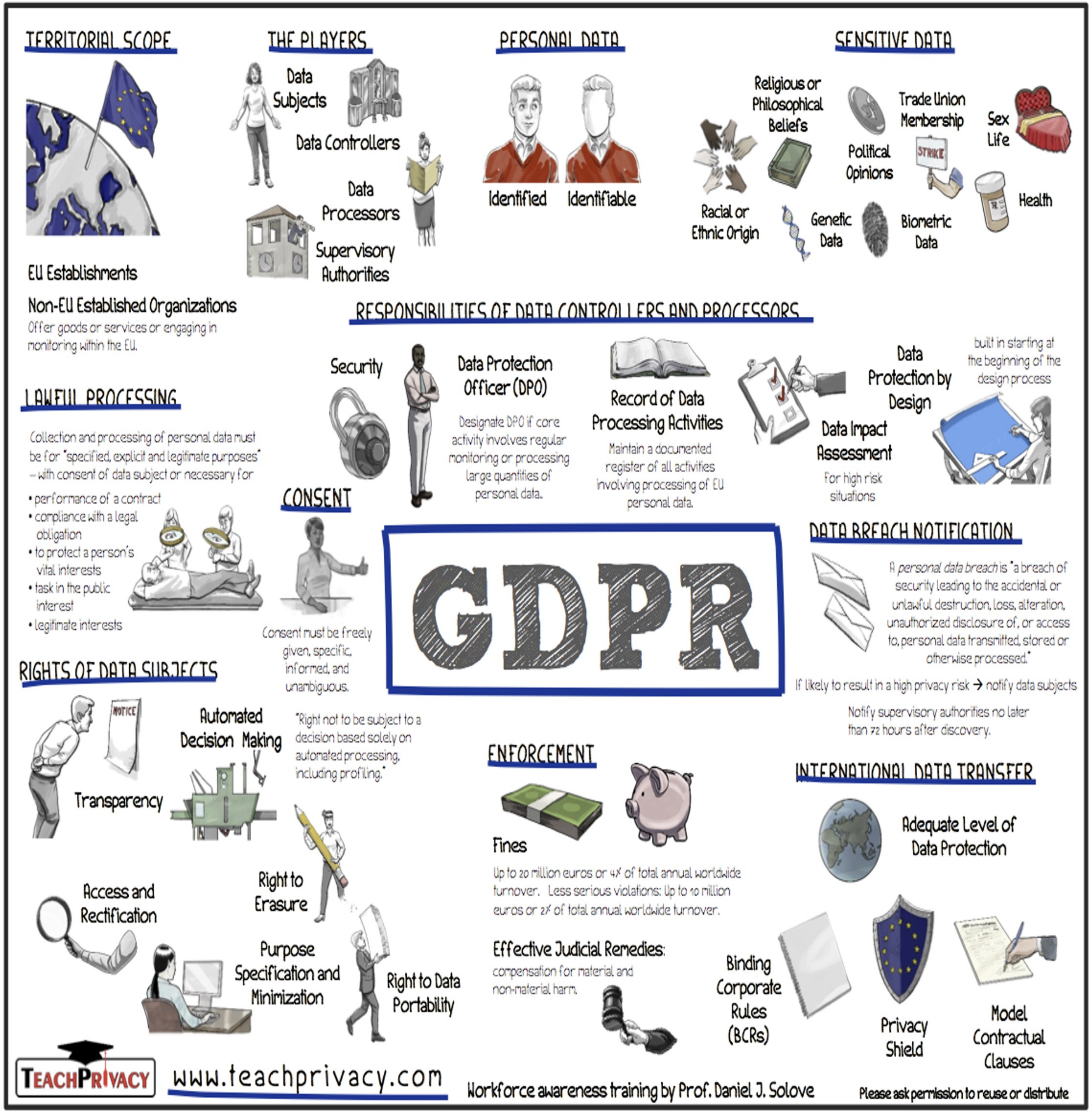
- Article 17 Right to erasure and “to be forgotten”
    - Paragraphs 1, 1a and 2a shall not apply to the extent that processing of the personal data is necessary: d. for archiving purposes in the public interest or for scientific, statistical and historical purposes in accordance with Article 89 (1)

Research exemption

- Article 89 “Safeguards and derogations for the processing of personal data for archiving purposes in the public interest, or scientific and historical research purposes or statistical purposes”
  - Appropriate safeguards to protect the right and freedoms of the data subject
  - Technical and organizational measures
  - The principle of data minimization
  - Pseudonymization if possible (anonymity = no problem)
- Member states may maintain or introduce further conditions, with regard to the processing of genetic data, biometric data or health data (Art. 9, 4.)

GDPR summary

- Painful but useful
- Elaboration on good practices for personal + sensitive data
- Increased accountability (cannot get away with “pretend”)
- Serious penalties possible
- GDPR does not forbid processing of personal data, but handle with care (consent, transparency, minimization, etc)
- Consent is not the only ground for processing
- Individual rights intended to protect personal autonomy
- Profiling and automated decision making require extra care
- Research exemption could open serious loopholes
- Research exemption makes secondary use of personal data possible, but handle with care



GDPR

TERRITORIAL SCOPE

EU Establishments

Non-EU Established Organizations

Offer goods or services or engaging in monitoring within the EU.

THE PLAYERS

Data Subjects

Data Controllers

Data Processors

Supervisory Authorities

PERSONAL DATA

Identified

Identifiable

SENSITIVE DATA

Religious or Philosophical Beliefs

Political Opinions

Trade Union Membership

Sex Life

Health

Racial or Ethnic Origin

Genetic Data

Biometric Data

RESPONSIBILITIES OF DATA CONTROLLERS AND PROCESSORS

Security

Data Protection Officer (DPO)

Designate DPO if core activity involves regular monitoring or processing large quantities of personal data.

Record of Data Processing Activities

Maintain a documented register of all activities involving processing of EU personal data.

Data Protection by Design

built in starting at the beginning of the design process

Data Impact Assessment

For high risk situations

LAWFUL PROCESSING

Collection and processing of personal data must be for "specified, explicit and legitimate purposes" – with consent of data subject or necessary for

- performance of a contract
- compliance with a legal obligation
- to protect a person's vital interests
- task in the public interest
- legitimate interests

CONSENT

Consent must be freely given, specific, informed, and unambiguous.

DATA BREACH NOTIFICATION

A *personal data breach* is "a breach of security leading to the accidental or unlawful destruction, loss, alteration, unauthorized disclosure of, or access to, personal data transmitted, stored or otherwise processed."

If likely to result in a high privacy risk → notify data subjects

Notify supervisory authorities no later than 72 hours after discovery.

RIGHTS OF DATA SUBJECTS

Transparency

Automated Decision Making

"Right not to be subject to a decision based solely on automated processing, including profiling."

Right to Erasure

Access and Rectification

Purpose Specification and Minimization

Right to Data Portability

ENFORCEMENT

Fines

Up to 20 million euros or 4% of total annual worldwide turnover. Less serious violations: Up to 10 million euros or 2% of total annual worldwide turnover.

Effective Judicial Remedies:

compensation for material and non-material harm.

INTERNATIONAL DATA TRANSFER

Adequate Level of Data Protection

Binding Corporate Rules (BCRs)

Privacy Shield

Model Contractual Clauses

TEACHPRIVACY www.teachprivacy.com

Workforce awareness training by Prof. Daniel J. Solove

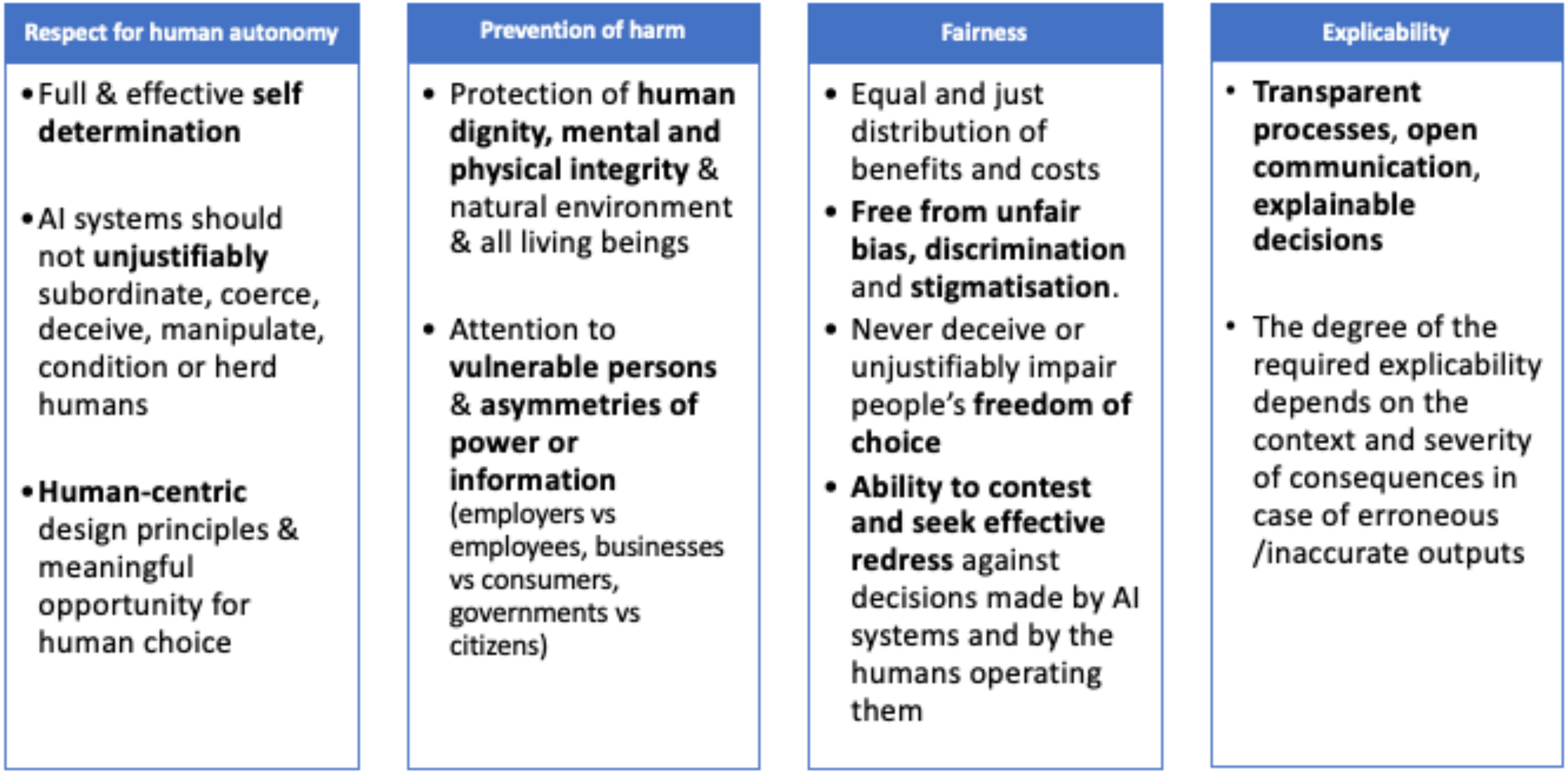
Please ask permission to reuse or distribute

Trustworthy AI

Trustworthy AI

- Trustworthy AI: lawful, ethical, robust (= will not cause any unintentional harm)
  - 4 ethical principles
    - Respect for human autonomy
    - Prevention of harm
    - Fairness
    - Explainability

Trustworthy AI



Respect for human autonomy	Prevention of harm	Fairness	Explicability
• Full & effective **self determination** • AI systems should not **unjustifiably** subordinate, coerce, deceive, manipulate, condition or herd humans • **Human-centric** design principles & meaningful opportunity for human choice	• Protection of **human dignity, mental and physical integrity** & natural environment & all living beings • Attention to **vulnerable persons & asymmetries of power or information** (employers vs employees, businesses vs consumers, governments vs citizens)	• Equal and just distribution of benefits and costs • **Free from unfair bias, discrimination** and **stigmatisation**. • Never deceive or unjustifiably impair people's **freedom of choice** • **Ability to contest and seek effective redress** against decisions made by AI systems and by the humans operating them	• **Transparent processes, open communication, explainable decisions** • The degree of the required explicability depends on the context and severity of consequences in case of erroneous /inaccurate outputs

Trustworthy AI

- 7 key requirements to be taken into consideration for Trustworthy AI

Human agency and oversight

Technical robustness and safety

Privacy and data governance

Transparency

Diversity, non-discrimination and fairness

Environmental and societal well-being

Accountability

Trustworthy AI

Risk-based approach – high risk / low risk

- Extra requirements for high-risk AI

2 cumulative criteria to determine high-risk AI

- **The AI application is employed in a high-risk sector**
  - ✓Sectors to be **specifically and exhaustively listed** in the new regulatory framework

    E.g., healthcare; transport; energy and parts of the public sector
  - ✓List to be periodically reviewed and amended, where necessary
- **The way the AI application is used is riskier**
  - ✓Not every use of AI in the selected sectors necessarily involves significant risks
  - ✓Assessment of the level of risk of a given use could be based on the impact on the affected parties

    E.g., AI applications that produce legal or similarly significant effects for the rights of an individual or a company; that pose risk of injury, death or significant material or immaterial damage; that produce effects that cannot reasonably be avoided by individuals or legal entities

Trustworthy AI

Always high-risk AI, regardless of the criteria:

- AI applications used for recruitment & in situations impacting workers rights
- AI applications for the purposes of remote biometric identification and other intrusive surveillance technologies

Key features of the requirements for high-risk AI

- training data
- data and record-keeping
- information to be provided
- robustness and accuracy
- human oversight
- specific requirements for certain AI applications, such as those used for purposes of remote biometric identification

Trustworthy AI

- More granular/nuanced risk-based approach
  - Unacceptable Risk – Prohibited AI Practices (next slide)
  - High-risk AI systems
    - AI systems intended to be used as safety component of products & subject to third party ex-ante conformity assessment
    - Other AI systems with mainly fundamental rights implications
  - Low or minimal risk

Trustworthy AI

Unacceptable Risk - Prohibited AI Practices

- Practices **contravening Union values** (e.g. violating fundamental rights)
- Practices likely to **manipulate persons** or **exploit vulnerabilities of vulnerable groups** (e.g. children, persons with disabilities) to materially distort their behaviour, likely to cause psychological/physical harm
- **AI-based social scoring** for general purposes done by public authorities
- The use of **'real time' remote biometric identification systems in publicly accessible spaces** for the purpose of **law enforcement**

3 Exceptions, subject to prior authorisation:

- **targeted search for specific potential victims** of crime, including missing children
- **prevention of** a specific, substantial and imminent **threat to the life or physical safety** of natural persons **or** of a **terrorist attack**
- **detection, localisation, identification or prosecution** of a **perpetrator or suspect of a criminal offence**

Trustworthy AI

High-risk AI systems → Stricter requirements

Risk management system

Data and data governance

Technical documentation and recording keeping

Transparency and provision of information to users

Human oversight

Robustness

Accuracy and security

- AI Regulation Proposal Annex III – Article 6 provides a list of the high-risk AI systems used by Law Enforcement
- Additional obligations for providers, manufacturers, importers, distributors, users etc. of high-risk AI systems
- Self assessment & conformity assessment by 3rd parties in specific cases (Article 43)
- Exceptions for surveillance (next slide)

Trustworthy AI

Transparency obligations for certain AI systems

Natural persons shall be informed that they are interacting with an AI system	/!\ This obligation shall not apply to AI systems **authorised by law to detect, prevent, investigate and prosecute criminal offences**, unless those systems are available for the public to report a criminal offence
Users of an emotion recognition system or a biometric categorisation system shall inform of the operation of the system the natural persons exposed thereto	/!\ This obligation shall not apply to AI systems used for biometric categorisation, which are **permitted by law to detect, prevent and investigate criminal offences**
Users of an AI system that generates or manipulates 'deep fake' content shall disclose that the content has been artificially generated or manipulated	/!\ This obligation shall not apply where the use is **authorised by law to detect, prevent, investigate and prosecute criminal offences** or it is necessary for the exercise of the right to **freedom of expression** and the right to freedom of the arts and sciences

Artificial Intelligence Act

Aim: To (i) ensure AI systems placed on the EU market are safe and respect existing law, (ii) ensure legal certainty to facilitate investment and innovation in AI, (iii) enhance governance and effective enforcement; and (iv) facilitate the development of a single market for lawful, safe and trustworthy AI

Main purpose(s): To create harmonised rules for the development, placing on the market, and use of AI in the EU

Sectors mainly impacted: All sectors

Governance and Enforcement: European AI Board and National Competent Authorities (NCAs)

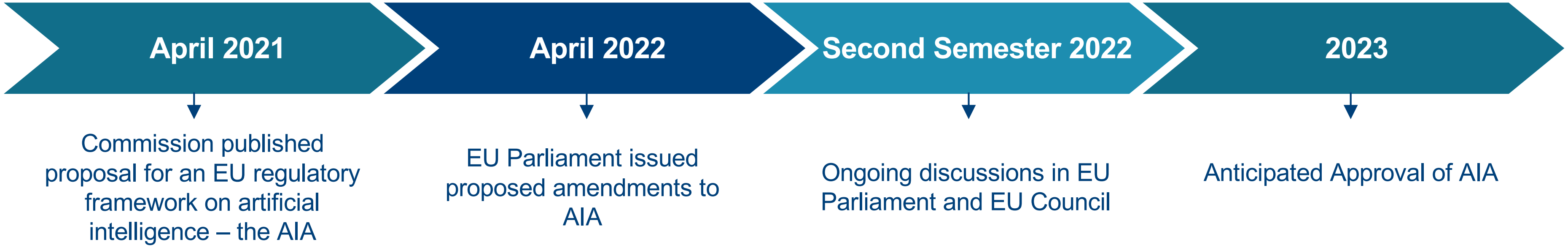
Maximum fines of 6% annual worldwide turnover or €30 m (i.e., higher than GDPR)

Practical impact: Applies primarily to providers (i.e., the entity that develops or has an AI system developed) and users of AI systems.

AIA seeks to regulate AI systems in accordance with the level of risk they present.

Extraterritorial scope: Applies primarily to: (i) Providers of AI systems placing AI systems on the EU market (irrespective of location of provider), (ii) Users of AI systems located in the EU, and (iii) Providers and users of AI systems located in third country where the output of those systems are used in the EU.

Expected Timeline



April 2021

Commission published proposal for an EU regulatory framework on artificial intelligence – the AIA

April 2022

EU Parliament issued proposed amendments to AIA

Second Semester 2022

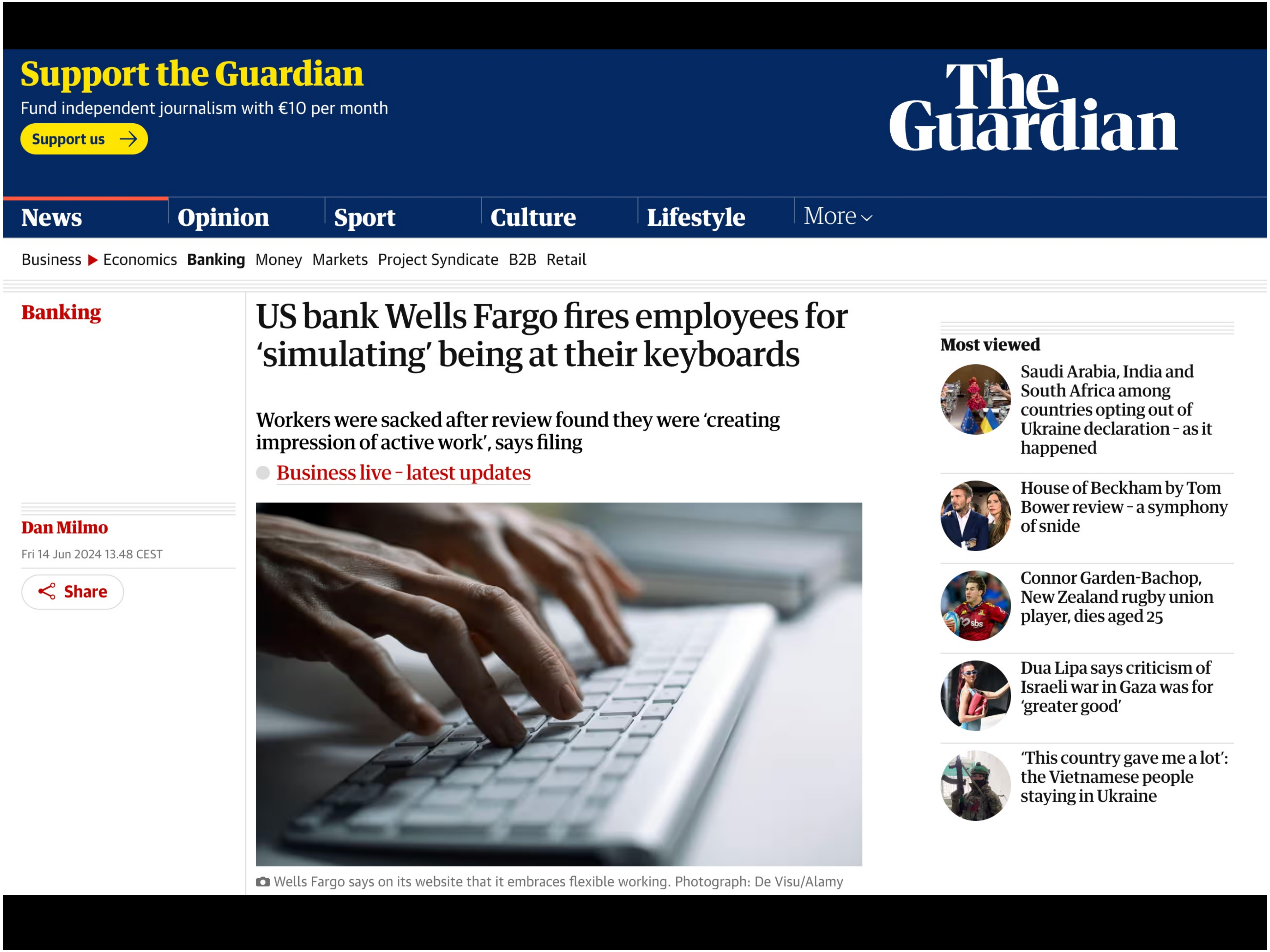
Ongoing discussions in EU Parliament and EU Council

2023

Anticipated Approval of AIA

Risk-Based Approach to AI

	Concept	Interpretation under AIA (*Commission draft)
Article 5	Unacceptable Risk	• Prohibited under the AIA because considered a threat to safety / rights of individuals. • Include: (i) social scoring by public authorities, (ii) exploitation of vulnerable groups in society or manipulation of behaviour using specific techniques, and (iii) real-time remote biometric identification systems used in publicly accessible spaces for law enforcement purposes.
Article 52	Limited Risk	• Subject to transparency requirements and include: (i) chatbots, (ii) deep fakes, and (iii) emotion recognition systems.
Article 69	Minimal Risk	• Unregulated but encouraged to voluntarily comply with the requirements under the AIA for high risk AI systems through a Code of Conduct. Examples include e.g., spam filters.
Article 6	High Risk	• (i) AI system is a safety component in a product or is a product itself protected under specific EU legislation identified in Annex II; and (ii) this product is required to undergo a conformity assessment pursuant to this EU legislation; OR • AI systems mentioned in Annex III are "high risk" by default (e.g., automated hiring, biometric ID systems).



The Guardian

News | Opinion | Sport | Culture | Lifestyle | More

Business ▶ Economics **Banking** Money Markets Project Syndicate B2B Retail

Banking

US bank Wells Fargo fires employees for 'simulating' being at their keyboards

Workers were sacked after review found they were 'creating impression of active work', says filing

● Business live - latest updates

Dan Milmo

Fri 14 Jun 2024 13.48 CEST

Share

Wells Fargo says on its website that it embraces flexible working. Photograph: De Visu/Alamy

Most viewed

- Saudi Arabia, India and South Africa among countries opting out of Ukraine declaration - as it happened
- House of Beckham by Tom Bower review - a symphony of snide
- Connor Garden-Bachop, New Zealand rugby union player, dies aged 25
- Dua Lipa says criticism of Israeli war in Gaza was for 'greater good'
- 'This country gave me a lot': the Vietnamese people staying in Ukraine