Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

What is the Internet Doing to Me:

Ethics on the Internet
Emerging Practice with AI and Data

Dave Lewis, dave.lewis@scss.tcd.ie
Thanks to: Wessel Reijers, Arturo Calvo, Killian Levacher

Student Online Teaching Advice Notice

The materials and content presented within this session are intended solely for use in a context of teaching and learning at Trinity.

Any session recorded for subsequent review is made available solely for the purpose of enhancing student learning.

Students should not edit or modify the recording in any way, nor disseminate it for use outside of a context of teaching and learning at Trinity.

Please be mindful of your physical environment and conscious of what may be captured by the device camera and microphone during videoconferencing calls.

Recorded materials will be handled in compliance with Trinity's statutory duties under the Universities Act, 1997 and in accordance with the University's policies and procedures.

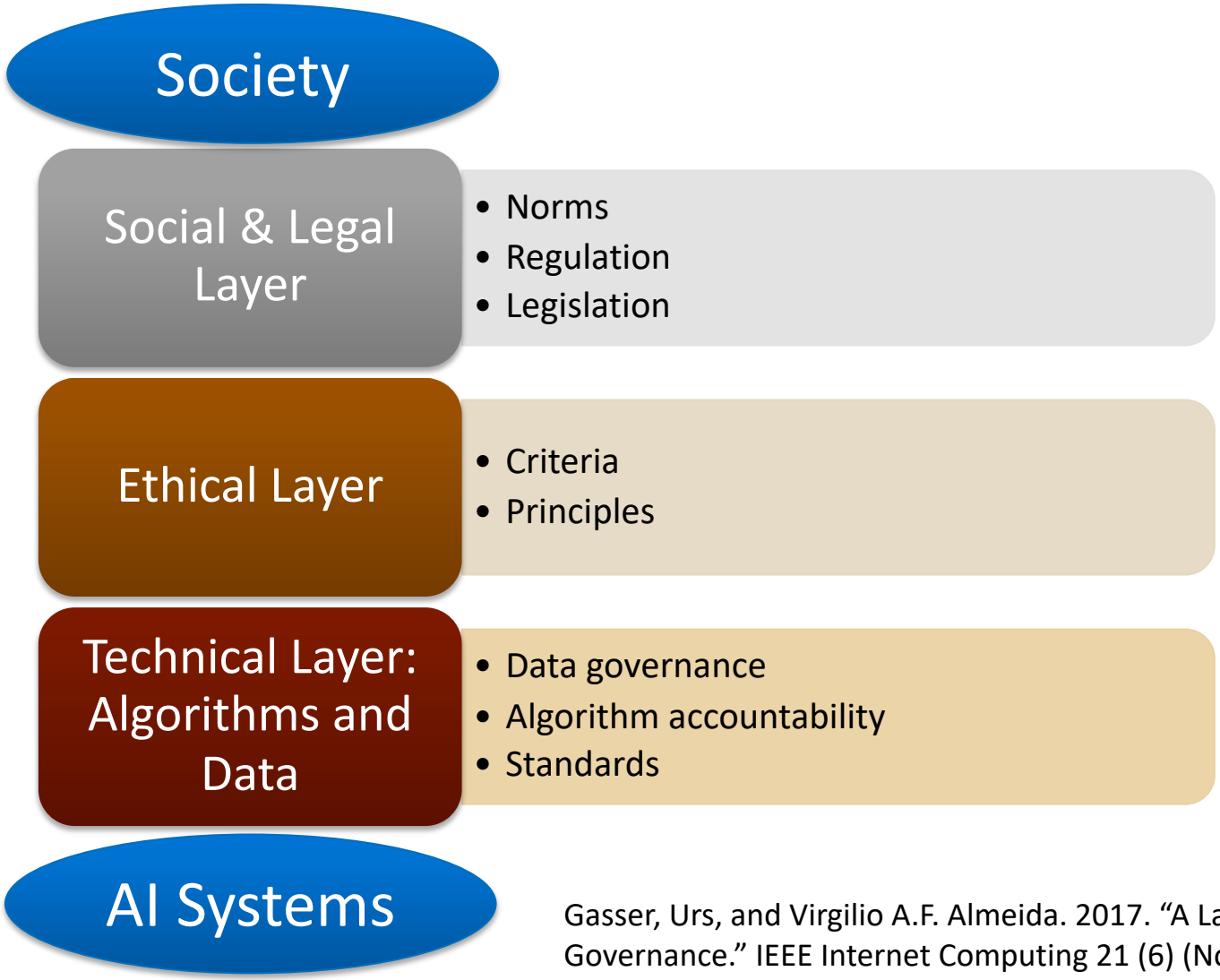
Further information on data protection and best practice when using videoconferencing software is available at https://www.tcd.ie/info_compliance/data-protection/.

AI Ethics – where next?

- Several governments, public bodies and companies have proposed trustworthy/ethical AI principles
- EU, USA and China now proposing legislation
- AI industry has failed to form trusted self-regulation – companies trying their own, e.g. FaceBook ‘Supreme Court’
- Guidance for industry emerging: proprietary training/consulting and in the form of industry standards.

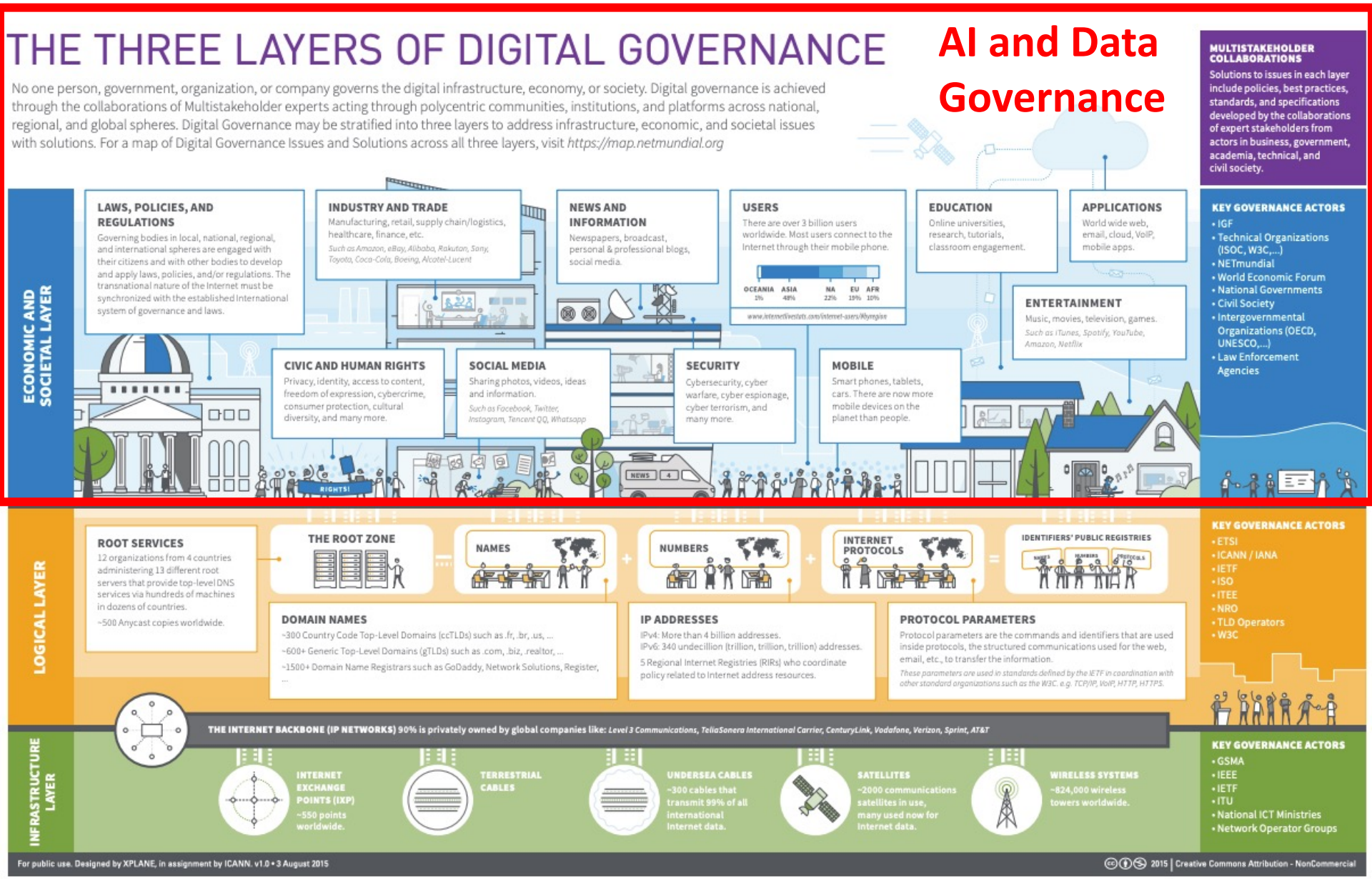
https://www.coe.int/en/web/artificial-intelligence/national-initiatives

AI Governance: Layered Model



Gasser, Urs, and Virgilio A.F. Almeida. 2017. "A Layered Model for AI Governance." IEEE Internet Computing 21 (6) (November): 58–62.

Governance over the Internet



Challenges in Governing Ethical AI and Data technology

- **Definition**: Difficult to reach stable consensus on what defines AI
- **Discreetness**: Growing access to AI skills and computing power, it can be developed out of sight
- **Diffuseness**: AI used in a diffuse set of locations and jurisdictions
- **Discreteness**: Impact of an AI component only apparent when assembled into a system
- **Opacity:** Modern machine learning yields results without clear explanations
- **Forseeability**: AI-driven autonomous system can behave in unforseeable ways – 'liability gap'
- **Control**: AI can work in ways/speeds out of control of those responsible for them

Scherer, M.U. Regulating Artificial Intelligence System, Harvard Journal of Law and Technology, 29(2) 2016

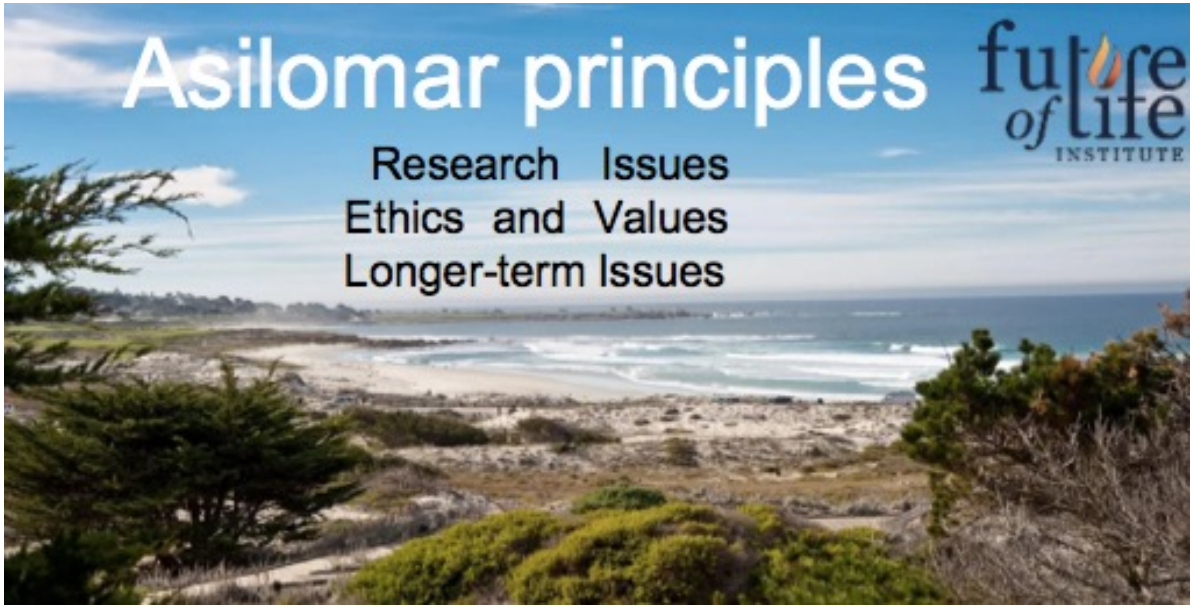
Headwinds to Consensus on AI Governance

- **Pacing:** AI tech and applications develop faster that societies' ability to regulate it
- **Securitisation**: International competition as AI perceived as a strategic economic/military resource
- **Innovation:** Perceived impediment to AI-based innovation and its economic and social benefits
- **Asymmetry:** Power of AI concentrated in a few digital platforms that benefit from massive network effects

Asilomar Principles

Ethical AI Principles

- Safety
- Failure Transparency
- Judicial Transparency
- Responsibility
- Value Alignment
- Human Values
- Personal Privacy
- Liberty and Privacy
- Shared Benefit
- Share Prosperity
- Human Control
- Non-subversion
- AI Arms Race



https://futureoflife.org/ai-principles/

Examples of Ethical Principles: EU Ethics Guidelines for Trustworthy AI - 2019

Ethical Principles mapped from EU Charter of Fundamental Right

International AI Policy Differentiator for EU

Ethical AI, alongside Lawful AI and Robust AI

Requirements/Principles

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

https://ec.europa.eu/futurium/en/ai-alliance-consultation

EU Ethics Guidelines for Trustworthy AI

Risk Mitigation Methods

- Technical:
  - Architecture,
  - Ethics/privacy-by-design,
  - Explanation,
  - Testing/validation,
  - QoS Indicators
- Non Technical:
  - Regulation
  - Code of Conduct
  - Standardisation
  - Certification
  - Accountability via Governance Frameworks
  - Education & Awareness
  - Stakeholder Participation
  - Diverse Design Teams

https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

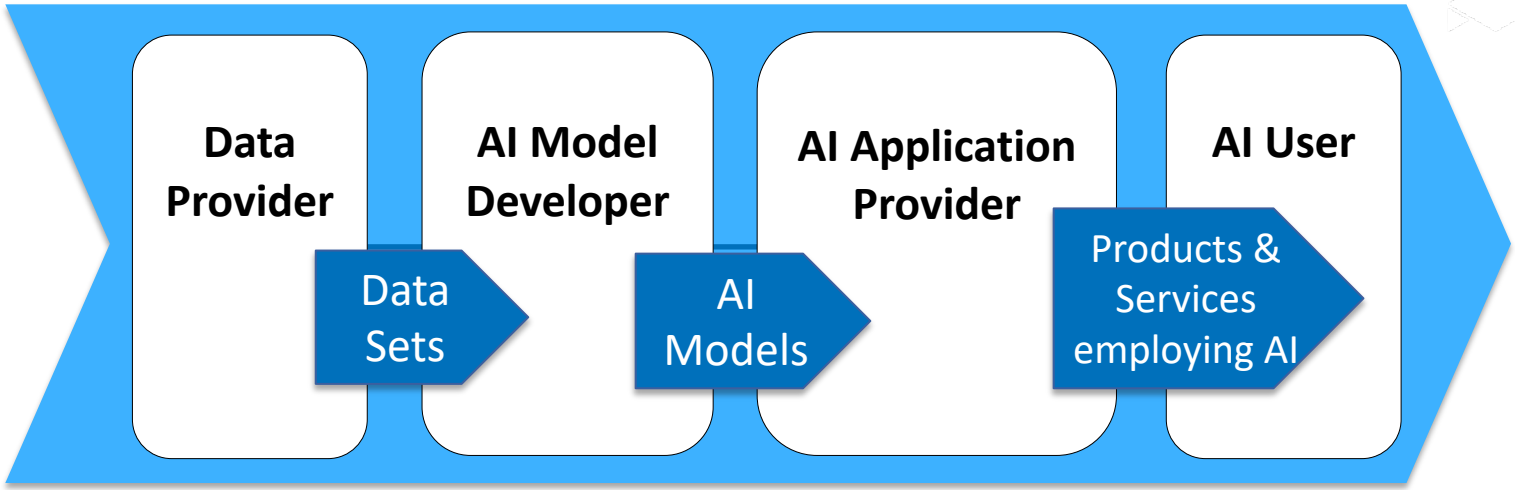
Competing/Converging Sets of Principles

Consensus on principles of

- Transparency
- Justice
- Non-maleficence
- Responsibility
- Privacy

Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. Nat Mach Intell 1, 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2

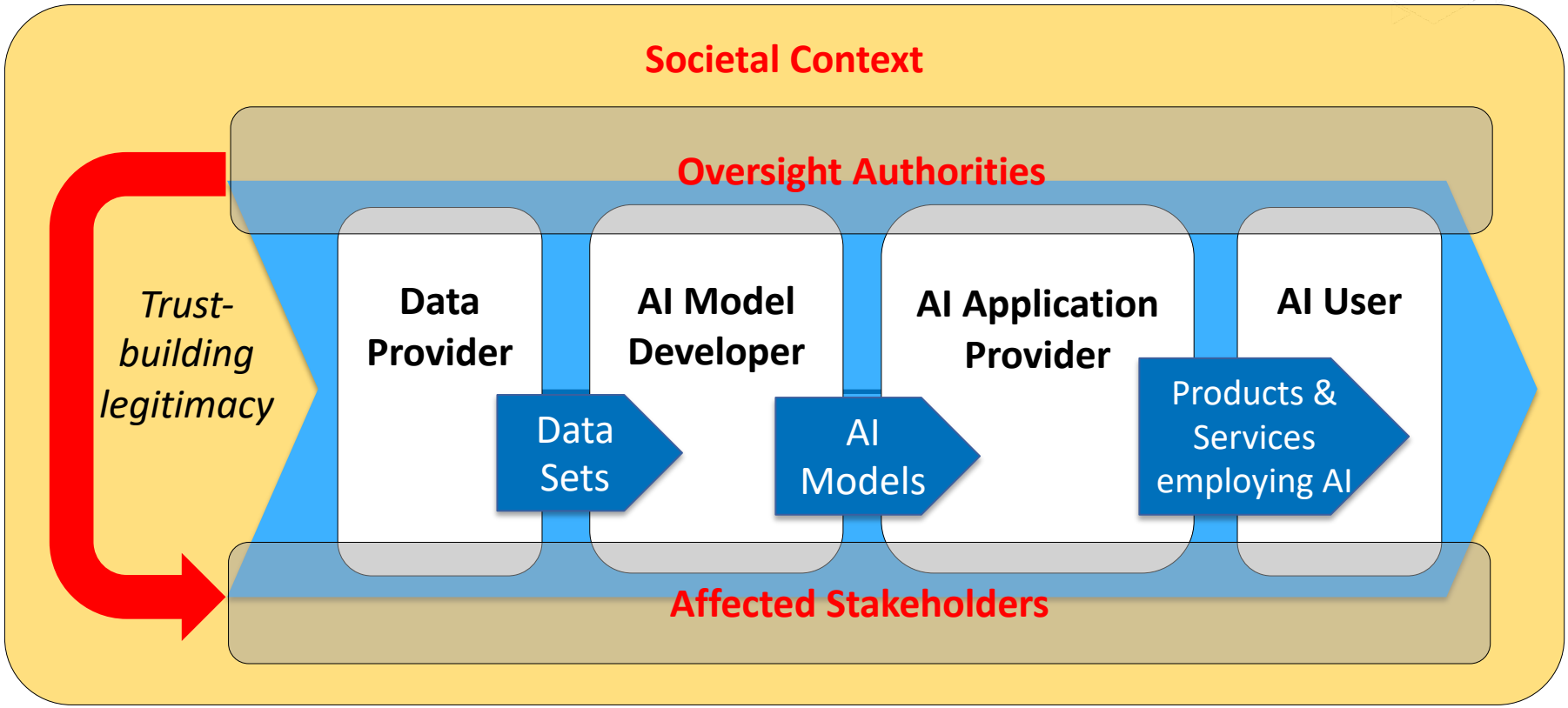
AI Value Chain Roles for Governance



- Data Provider: Official data sources, Data Brokers, You!
- AI Creator: Uses data to build/train AI models
- AI Operator: Uses AI models, perhaps several, in a product or service
- AI User: Decision makers, consumers, You!

- Organizations can take on several roles at once
- AI driven decisions can affect many people beyond User:
    - Your Friends/Family/Community, Patients, Students, Job or Insurance Applicants, Displaced Workers

Principles to Practice: Governance in context



Key Questions:

- Which (non-User) Stakeholders are affected? Your Friends/Family/Community, Patients, Students, Job or Insurance Applicants, Displaced Workers
- Who provides Oversight? Company, Ethics Panel, Sectoral Body, Government "FDA for Algorithms"?
- What Authority do they wield?
- Legitimacy of Oversight for affected Stakeholders?

Approaches to AI/Data Governance

- Self Governance by Organisations
  - Internal tech ethics boards – current examples lack transparency
- Self Regulation by Industries
  - Example: Partnership for AI: https://www.partnershiponai.org/
  - Lack of transparency and enforcement
- Government Regulation
  - EU white paper on AI: regulator "high risk" algorithms

    https://ec.europa.eu/digital-single-market/en/news/white-paper-artificial-intelligence-public-consultation-towards-european-approach-excellence
  - Labelling of AI projects akin to energy efficiency

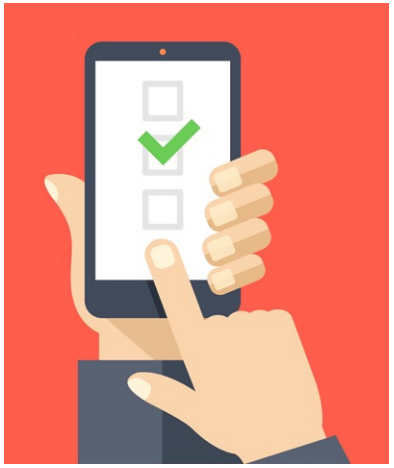
    Ethics much more complicated than energy consumption
- Hybrid:
  - Supplier declaration of conformance for data sets or trained models
  - External certification of declaration processes
- Machine Ethics
  - Stuart Russell: Human Compatible: AI and the Problem of Control

Does GDPR help in Governing AI/Data

- Informed consent, data minimization and storage limitations help
  - Data Subject Rights and Supervisory Authority offer some Transparency
  - Fines, SA judgement and compensation offers some Accountability
- Right to Portability: intended to give power to consumers, but where to port to?
- Pseudo anonymization: if personal information _can_ be extract from data sets it is subject to GDPR
  - As AI get better, more data sets are subject to regulation
  - Industry balance with statistical techniques, e.g. differential privacy
  - Profiling covered: inference of new data that “evaluates personal aspects”
- Right to Explanation and automated decision making:
  - Human explanation and intervention in automated decisions
  - Explainability of machine learning can make this a challenge

Limitations of GDPR

GDPR's **Notice and Consent** model limited by asymmetry in: Knowledge, Expertise & Time



Informed consent

Specific purpose, Data minimisation, Storage limitation

Sharing transparency

Rights to access, erase, correct, object, port

Cons

Pros

Vague but reassuring language "improve service"

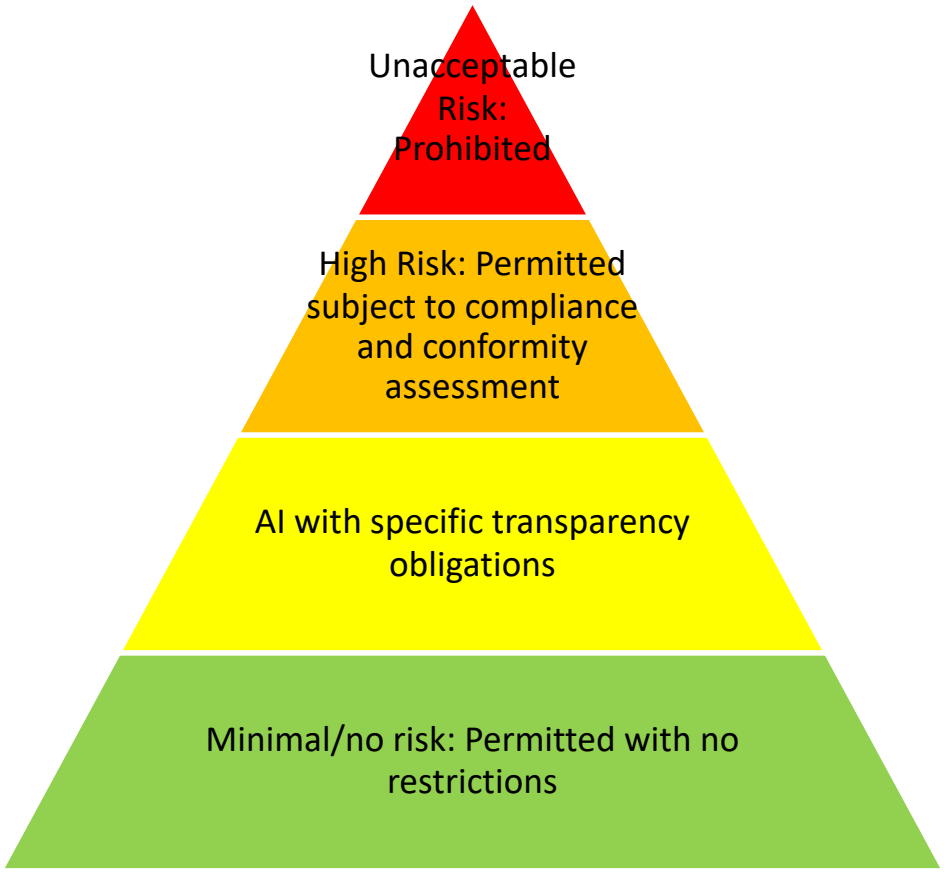
Dark Patterns for consent

"Legitimate Interest"

Poor enforcement

Data drives ever-more enticing and addictive apps

Proposed Framework for EU AI Act



Aims to ensure AI systems are **safe** and respect **fundamental rights**

Provide legal certainty for innovation, promote public trust and support single market

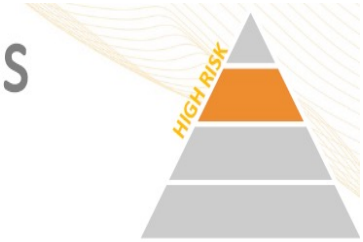
A **Risk-based** approach to regulating AI

Large penalties: 30M or 6% of global turn-over

https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence

AI Act: Transparency and Accountability

Requirements for high-risk AI systems (Title III, Chapter 2)

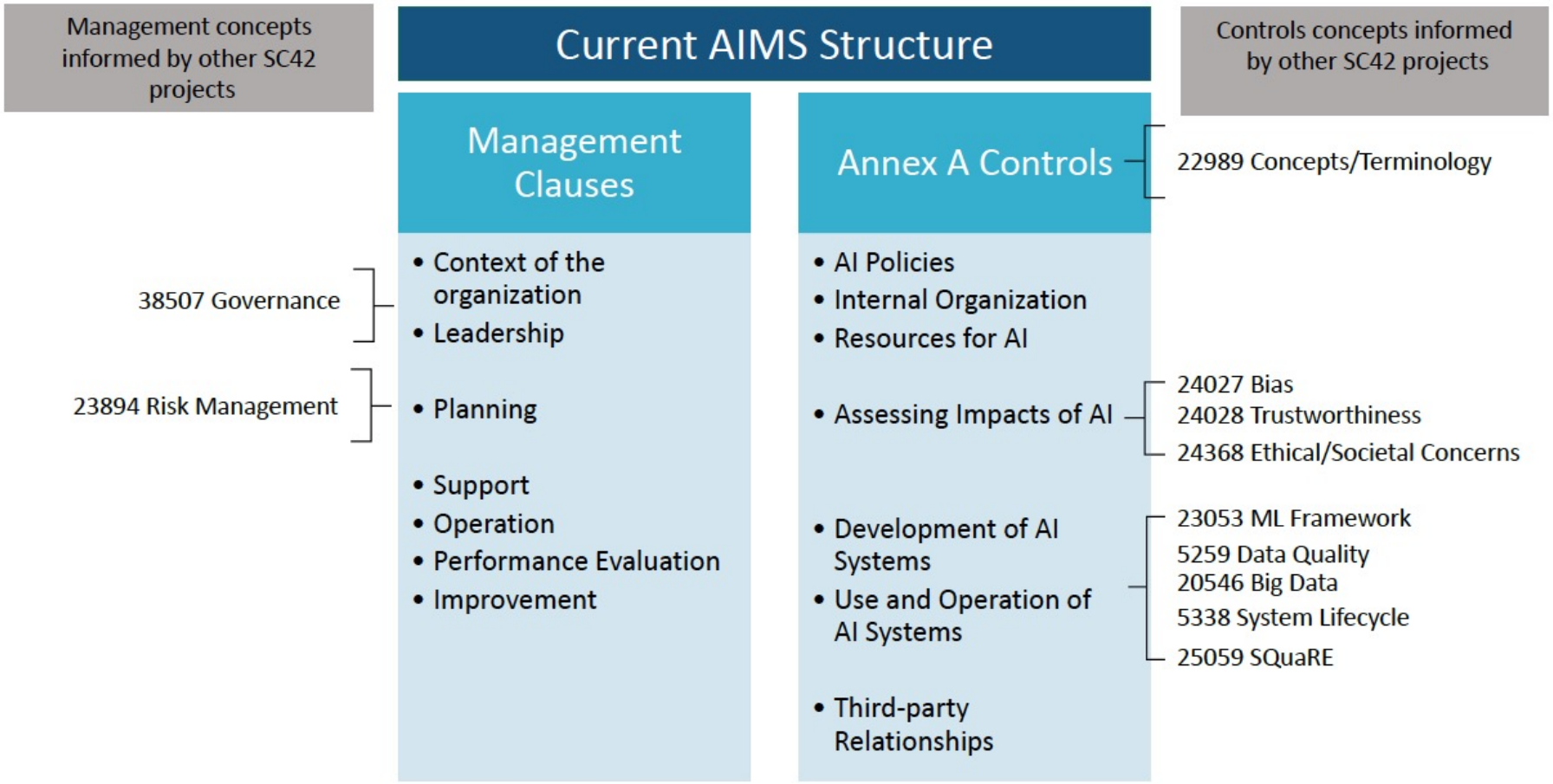


Establish and implement **risk management system** & in light of the **intended purpose** of the AI system

- Use high-quality **training, validation and testing data** (relevant, representative etc.)
- Draw up **technical documentation** & set up **logging capabilities** (traceability & auditability)
- Ensure appropriate degree of **transparency** and provide users with **information** on capabilities and limitations of the system & how to use it
- Ensure **human oversight** (measures built into the system and/or to be implemented by users)
- Ensure **robustness**, **accuracy** and **cybersecurity**

AI Management System

Industry Standards for Managing AI: SC42 AI Management System (AIMS) Proposal 42001



SC42: Social Responsibility for AI

- Ethical and Societal Issues:
  - ISO need international consensus BUT avoids importing specific value-sets
  - Needs principles, which ones?
- ISO already has non-ICT specific principles: ISO 26000 – Social Responsibility
- Identification and engagement with stakeholder is key

Principles	Core Subjects (stakeholders)
• Accountability • Transparency • Ethical behavior • Respect for stakeholder interests • Respect for the rule of law • Respect for international norms of behaviour • Respect for human rights	• Organizational Governance Mitigations (governance board, managers, shareholders) • Human Rights (everyone) • Labour Practices (workers) • The Environment (future generations) • Fair Operating Procedures (suppliers, customers, regulators) • Consumer Issues (consumers) • Community Involvement and Development (local communities)

Stakeholders: Human Rights

Risks

- **Legal**, from impacts in equality, privacy, access to justice
- **Reputational**, from impacts to dignity, physical and mental integrity
- **Complicity** in partner violations of rights
- **Conflicts between stakeholder**, e.g. investors vs consumers, suppliers vs local communities
- **To civil & political rights**: e.g. fake news social media bots, deep fake video impacting elections, filter bubbles, censorship
- **To economic, social, cultural rights**: education, healthcare, wellbeing
- **To just and favourable work**: casualised and deskilling labour of gig and click workers

Treatments

- *Due diligence*: Human rights policy
- *Avoid* value chain partners that may commit violations
- Establish *grievance and redress mechanism*: transparent, accessible, external scrutiny, e.g. AI explanations
- *Monitor for discrimination* towards vulnerable groups in AI decision making, e.g. insurance, justice, recruiting
- *Education and access* for all groups to benefit of AI
- Ensure worker *freedom of association* and collective bargaining

Stakeholders: Labour Practices - Workers

Risks	Treatments
• Legal: arising from **discrimination** in AI assisted recruiting • Reputational: **worker dissatisfaction**, e.g. to intrusive monitoring • AI automation leading to **labour displacement** • **Deskilling** of work, e.g. translators correct machine translations • To worker **physical and mental health**	• *Recognition* of secure employment & decent working conditions • Engage in *social dialo*gue with workers and affective professional and community representative • Employee *retraining* • *Health and safety practices*, e.g. for robot coworkers, offensive content moderators • Protect *personal data of employees* • Seek *assurance* of good labour practices in value chain partners

Stakeholders: The Environment – Future Generations

Risks

- Increased **carbon emission** due to AI training and service operation
- Resource usage and **pollution** from AI-driven product creation and disposal, e.g. sensors, batteries, obsolete smart phones

Treatments

- Monitor and plan reduction of non sustainable energy and resources use over whole product lifecycle
- Make AI services available for environmental monitoring and analysis

Stakeholders:
Fair Operating Procedures: Suppliers, Customers, Regulators

Risks	Treatments
• Use of AI in **corrupt or anti-competitive practices**, e.g. finance, investment, procurement • Use of AI to **undermine the public political process**, e.g. through deep fakes, targeted manipulation or misinformation online • Violation of **intellectual property** rights	• Ensure *transparency and other safeguards* against abuse of power or complicity, e.g. protecting whistleblowers • Promote responsible behaviour in *value chain partners*, e.g. through requesting ethical impact assessment from AI partners • Identify, respect and fairly compensate *right holders*, e.g. annotators, translators, content providers

Stakeholders: Consumers

Risks	Treatments
• Conveying **deceptive, misleading, fraudulent or unfair** information to consumers • Endangering **consumer health** and safety, e.g. mental health, self image • **Unsustainable consumption** • Misuse of **personal data** • Denial of access to **essential services**	• Clearly *identify promoted content* and its sponsors • Monitor and benchmark *safety performance* and correct problems promptly • Consumer '*nutrition labels*', e.g. performance and failure envelope, energy usage • Clear and accessible *complaint and redress* mechanisms • Compliance with *privacy regulations*, e.g. transparent on data held or shared and its use • Consumer *education and awareness* raising, regardless of their capabilities or accessibility needs • *No discrimination or censorship* in access to services and information

Stakeholders:
Community Involvement and Development

Risks	Treatments
• Difficulty **identifying communities** suffering negative impact of AI use (including non-users), e.g. social networks, road users, medical patients • Negative impact on **local health, employment and wellbeing**, e.g. deskilling, child development, culture wars • **Concentration of AI's wealth and income creation** away from local communities • Perpetuating **local dependence on philanthropic** activities	• *Consult* with early and *widely* with communities, especially where vulnerable • Be *transparent* on engagement with local authorities • Promote *education and preservation* of local cultures • Support *employment creation and skills development* in impacted communities and along value chain • Direct AI to *solve local* social and environmental issues • Enhance *local scientific and technological* development and entrepreneurship • Promote *economic diversification*, support local suppliers and employment

Role of **Data** in Trustworthy AI and Ethics

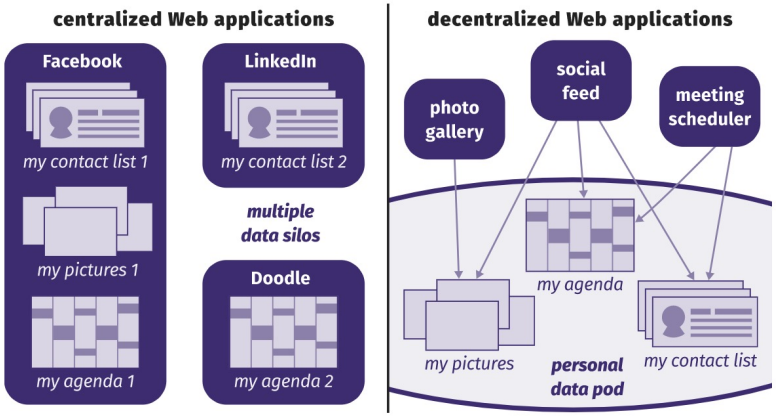
- Problem: for Data, **Possession** is 9.9/10^{ths} of the Law
- Power of AI-driven digital engagement (and then it potential power over us) grows with the volume (and quality) of its training data
- **Controlling the Flow of Data** is the Key to Governing AI

Data Governance

- Regulation can improve Transparency and Accountability of data handling in organisation, but can it do enough to maintain legitimacy?
- Regulations are highly technical and suffer pacing problem – how can Democratic oversight and control be exercised?
- Concentration of power over data, how can this be decentralised?

Personal Online DataStores

- Edge Platforms emerging for **maintaining possession** of Personal Data:
  - Inrupt.com (based on solid.org)
  - hubofallthings.com
- Software to keep your data in a store you control - “personal data pod”
- Companies that want data have to agree to **your** T&Cs rather than you agreeing to theirs



Pros:

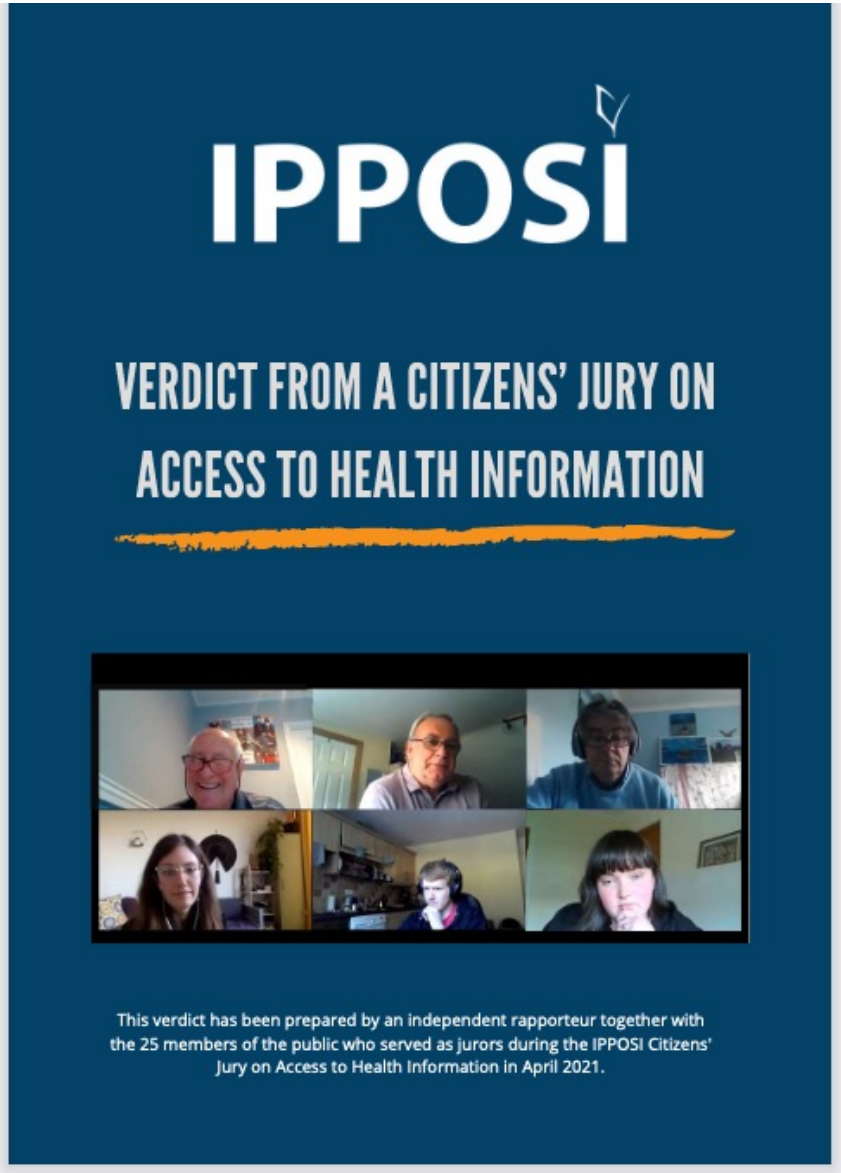
- Can consistently enforce your preferences for sharing data
- Can more readily rescind/renegotiate access
- Pool understanding of requests for your data
- No centralized data store to attract hackers

Cons:

- Requires some effort to monitor and control own data and set T&Cs
- Trustworthiness and legal basis of “Pod” provider – EU Data Governance Act - certified trusted intermediary

Mini-Publics to Deliberate Data and AI Issues

- Recent ‘Citizen Jury’ on Health Information
- National AI Strategy promises Youth Citizen Assembly on AI and an AI Ambassador



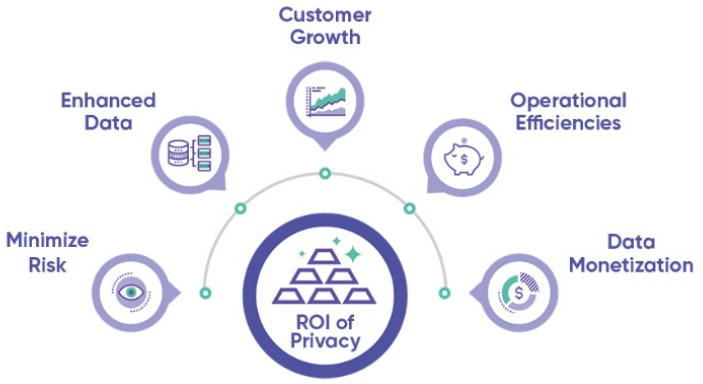
Could new patterns of Data Stewardship Help?

Organizations could already transfer governance responsibility to more representative groups:

- Data Unions – Data as Labour – https://blog.singularitynet.io
- Data Trusts
- Data Co-ops

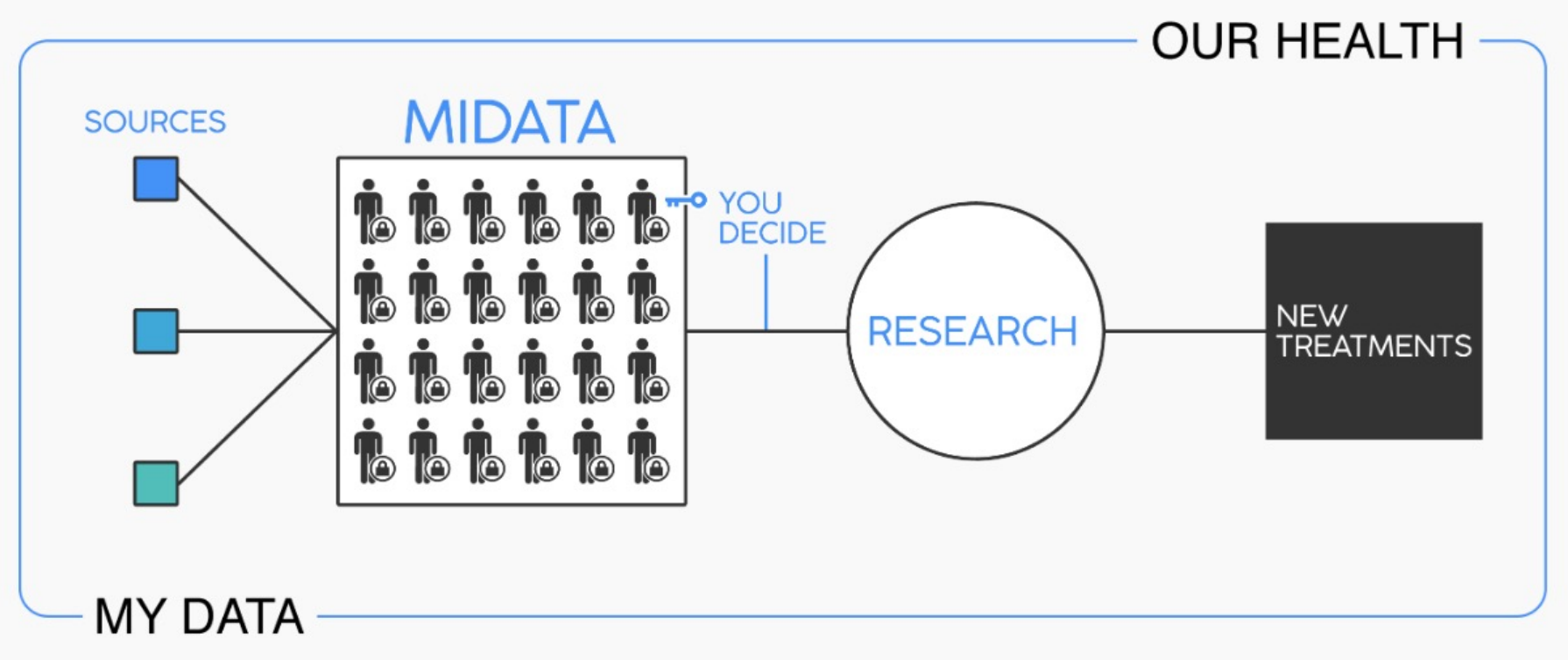
Example: Data Trusts

- Truata.com: Anonymised Data Analytics Services



- Offers large clients secure, anonymised analytics services of their own data
- Outsources data protection risks without loss of benefits from data analytics
- Part of business model is a Data Trust - constituted separately to the business/profit driven part of the company
- Data Trust gives clients (and their customers) confidence that the rules can't change for business reasons
- Other examples: https://theodi.org/article/odi-data-trusts-report/

Example: MIDATA Medical Data Coop



- Control of Data from hospitals and medical studies handed to MIDATA
- Operates as a **cooperative** in the interest of its members medical data subjects
- Management appointed and operated under **democratic principles**

https://www.midata.coop

Conclusions

- As tech becomes more powerful and ubiquitous, risks of individual and societal impact and harm grows
- Tech Ethics becoming a priority for governments and companies, e.g. for AI, Big Data, Robotics, IoT etc
- Modern innovation techniques feeding AI and Big Data applications need appropriate forms of ethical consideration – agile, accessible
- Ethic Canvas is a simple tool to help innovation teams reflect on ethical issues across application design iterations

References

- Principled Artificial Intelligence: Mapping Consensus in Ethical and Right-based Approaches to Principles for AI, Jessica Fjeld, Adamn Nagy, Berkman Klein Centre, Jan 15, 2020, https://cyber.harvard.edu/publication/2020/principled-ai
- AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, Floridi, L., Cowls, J., Beltrametti, M. et al. Minds & Machines (2018) 28: 689. https://doi.org/10.1007/s11023-018-9482-5
- Gasser, Urs, and Virgilio A.F. Almeida. 2017. “A Layered Model for AI Governance.” IEEE Internet Computing 21 (6) (November): 58–62.
- Hagendorff, T., (2019) "The Ethics of AI Ethics – An Evaluation of Guidelines", https://arxiv.org/pdf/1903.03425.pdf
- Adamson, G., Havens, J. C., & Chatila, R. (2019). "Designing a Value-Driven Future for Ethical Autonomous and Intelligent Systems". Proceedings of the IEEE, 107(3), 518–525. https://doi.org/10.1109/JPROC.2018.2884923
- Leenders, G. (2019). "The Regulation of Artificial Intelligence — A Case Study of the Partnership on AI". Medium, April, Retrieved from: https://becominghuman.ai/the-regulation-of-artificial-intelligence-a-case-study-of-the-partnership-on-ai-c1c22526c19f
- EU High Level Expert Group on AI, (2019). "Ethics Guidelines for Trustworthy AI", April 2019, Retrieved from https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines
- "Mapping regulatory proposals for artificial intelligence in Europe", Access Now, Nov 2018. Retrieved from https://www.accessnow.org/mapping-regulatory-proposals-AI-in-EU
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé, H., Crawford, K. (2018) "Data sheets for Datasets", in Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning, Stockholm, Sweden, 2018, available at: https://arxiv.org/abs/1803.09010
- Buchanan, B., Miller, T. (2017) "Machine Learning for Policy Makers - What it is and Why it matters" Belfer Centre, available from: https://www.belfercenter.org/sites/default/files/files/publication/MachineLearningforPolicymakers.pdf
- Joshua A. Kroll , Joanna Huey , Solon Barocas , Edward W. Felten , Joel R. Reidenberg , David G. Robinson & Harlan Yu, "Accountable Algorithms", 165 U. Pa. L. Rev. 633 (2017). Available at: https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3
- Calo, Ryan, "Artificial Intelligence Policy: A Primer and Roadmap" (August 8, 2017). Available at SSRN: https://ssrn.com/abstract=3015350 or http://dx.doi.org/10.2139/ssrn.3015350
- Future of Life Institute. (2017). "Asilomar AI Principles". Future of Life Institute. Retrieved from https://futureoflife.org/ai-principles/
- Scherer, Matthew U., "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies" (May 30, 2015). Harvard Journal of Law & Technology, Vol. 29, No. 2, Spring 2016. Available at SSRN: https://ssrn.com/abstract=2609777 or http://dx.doi.org/10.2139/ssrn.2609777
- Saurwein, Florian and Just, Natascha and Latzer, Michael, "Governance of Algorithms: Options and Limitations" (July 14, 2015). info, Vol. 17 No. 6, pp. 35-49, 2015. Available at SSRN: https://ssrn.com/abstract=2710400

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Thank You