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

Ethics in Technology Innovation and the Ethics Canvas

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.

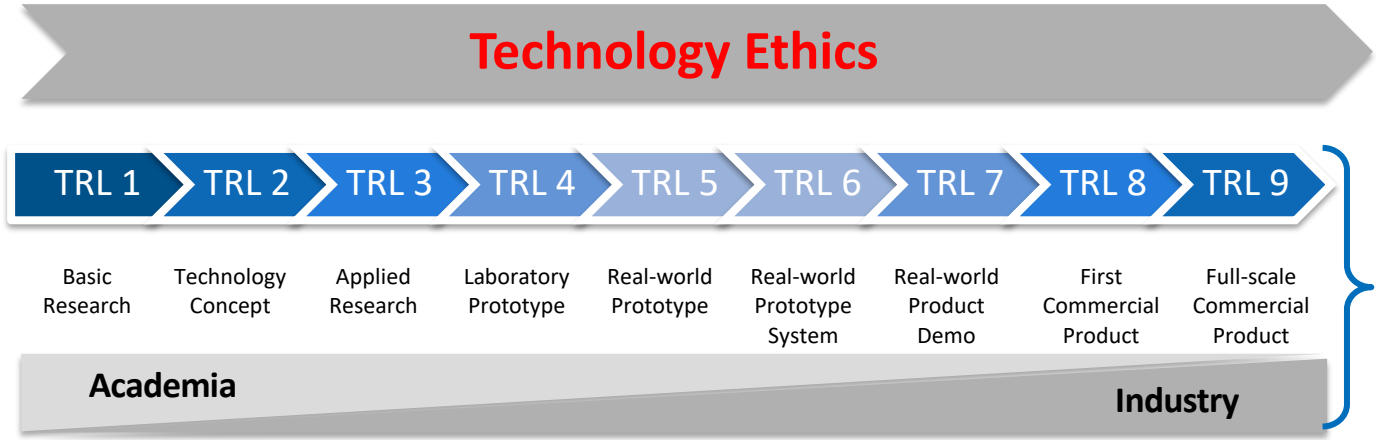
Further information on data protection and best practice when using videoconferencing software is available at https://www.tcd.ie/info_compliance/data-protection/.

Why Should Tech Innovators be Concerned with Ethics?

- Because new technologies have a **profound impact** on the way **we live**, on the **relationships we have**, on the **societal & political processes we engage in**.

- For tech innovators?
  - First: because it is good for the image of your business (instrumental goal)
  - Second: because it actually improves the service you provide! (substantive goal)
  - Third: because it is the *good* thing to do, it contributes to your idea of a better society and being a good person (normative goal)

Technology Ethics in Context



Technology Ethics concerns how technology impact society – impact from basic research to market deployment

Research Ethics

Research Ethics specifically addresses how research can be conducted without harming human participants

GDPR

Data protection applies at all stages of technology R&D

Ethics for publication, e.g. Neurips

AI Regulation

AI and Data bring a new initiatives on ICT R&D ethics

Dominant Views in Technology Ethics

- The neutrality thesis: technologies are *instruments* that we can use to attain our own goals.
  - "People kill people"
- The determinism thesis: technologies *dictate* everything we do, they determine who we are.
  - "Guns kill people"
- The co-shaping thesis: technologies and humans together "construct" our social world.
  - "Gun-men kill people"

Technology Impact: Example

Software Malfeasance: Example

Unintended Impacts - Example: Gender in Google Translate

- Some languages, like Turkish, don't has gender specific pronouns
- Google translate has to guess the gender when translating in English
- Statements allocating gender to role reveal gender bias
- What is the source of this?
- Is it a problem?

Sample Google Translate output:

he is a soldier
she's a teacher
he is a doctor
she is a nurse

https://qz.com/1141122/google-translates-gender-bias-pairs-he-with-hardworking-and-she-with-lazy-and-other-examples/

Power of Big Data: Example: Cambridge Analytica

- Academic research into Psychographics (U. Cambridge) revealed the link between psychological profiles and Facebook profiles
- Correlated major psychological types to elements in the social graph: Openness, Conscientiousness, Extroversion, Agreeableness and Neuroticism
- Cambridge Analytica applied psychographics to help target political ads in 2016 US elections….

https://www.theguardian.com/news/2018/mar/17/data-war-whistleblower-christopher-wylie-faceook-nix-bannon-trump

Algorithmic Power on Behaviour & Worldview

- “Race to the Bottom … of the Brain Stem” Tristian Harris
- 70% of YouTube views are based on algorithmic recommendations
- Business model maximises video views to maximise ad views
- Outrage/fear/anger the most reliable reactions that drive us to keep watching
- -> Recommender algorithm inevitably drive us to content that builds outrage to keep us watching

- Evidence to US Congress: https://www.youtube.com/watch?v=WQMuxNiYoz4
- Agenda: https://humanetech.com/wp-content/uploads/2019/06/Technology-is-Downgrading-Humanity-Let%E2%80%99s-Reverse-That-Trend-Now-1.pdf

AI & Data is Mainstreaming Technology Ethics

- Companies harvest and utilise personal data on a **massive scale**
- Growing concerns about the **collection, linking, use and leakage** of personal data from **mobile devices, bio-sensors, cameras, GPS trackers and social media**.
- Machine Learning deliver new levels of **insights and predictions** about an individual's behaviour and also feeds increasingly **personalised AI-driven interactive digital experiences - Ads to Alexa**
- Individuals and groups **struggle to understand** the impact of personal information processing
- Companies, especially SMEs, often lack the knowledge and expertise needed to address these **complex legal and ethical issues**.

.

Example of Risks: Algorithmic selection of digital content

- Manipulation of individuals or groups,
- Diminishing variety that creates biased views and distortion of reality,
- Constraints on communication and freedom of expression,
- Threats to privacy and data protection rights,
- Social discrimination,
- Violation of intellectual property rights,
- Impact on the human brain and cognitive capacity and
- Algorithmic power over human behavior and development.

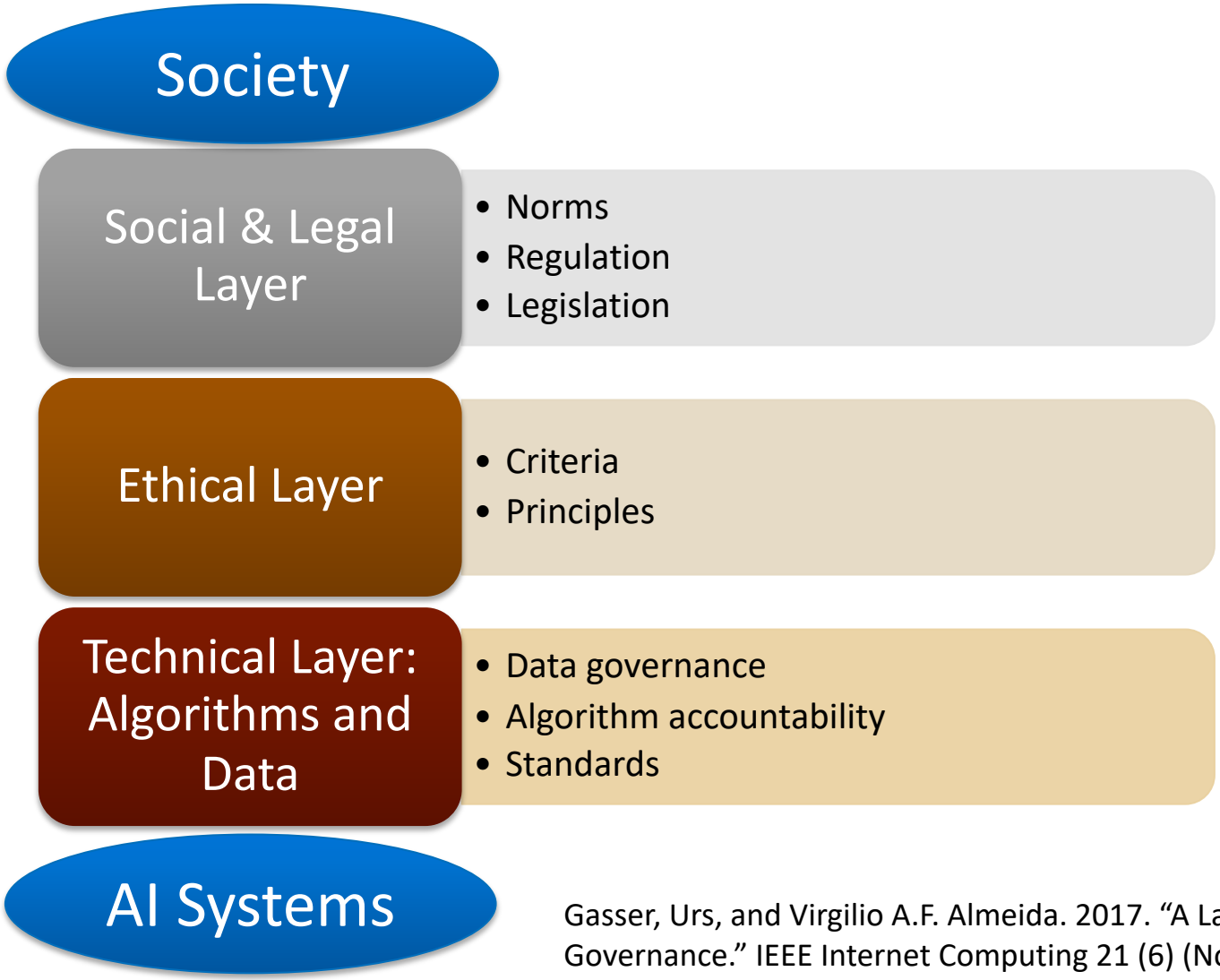
Latzer, M., Hollnbuchner, K., Just, N., & Saurwein, F. (2016). The economics of algorithmic selection on the Internet. Handbook on the Economics of the Internet, (October 2014), pp 395–425. Retrieved from https://doi.org/10.4337/9780857939852.00028

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.

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

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

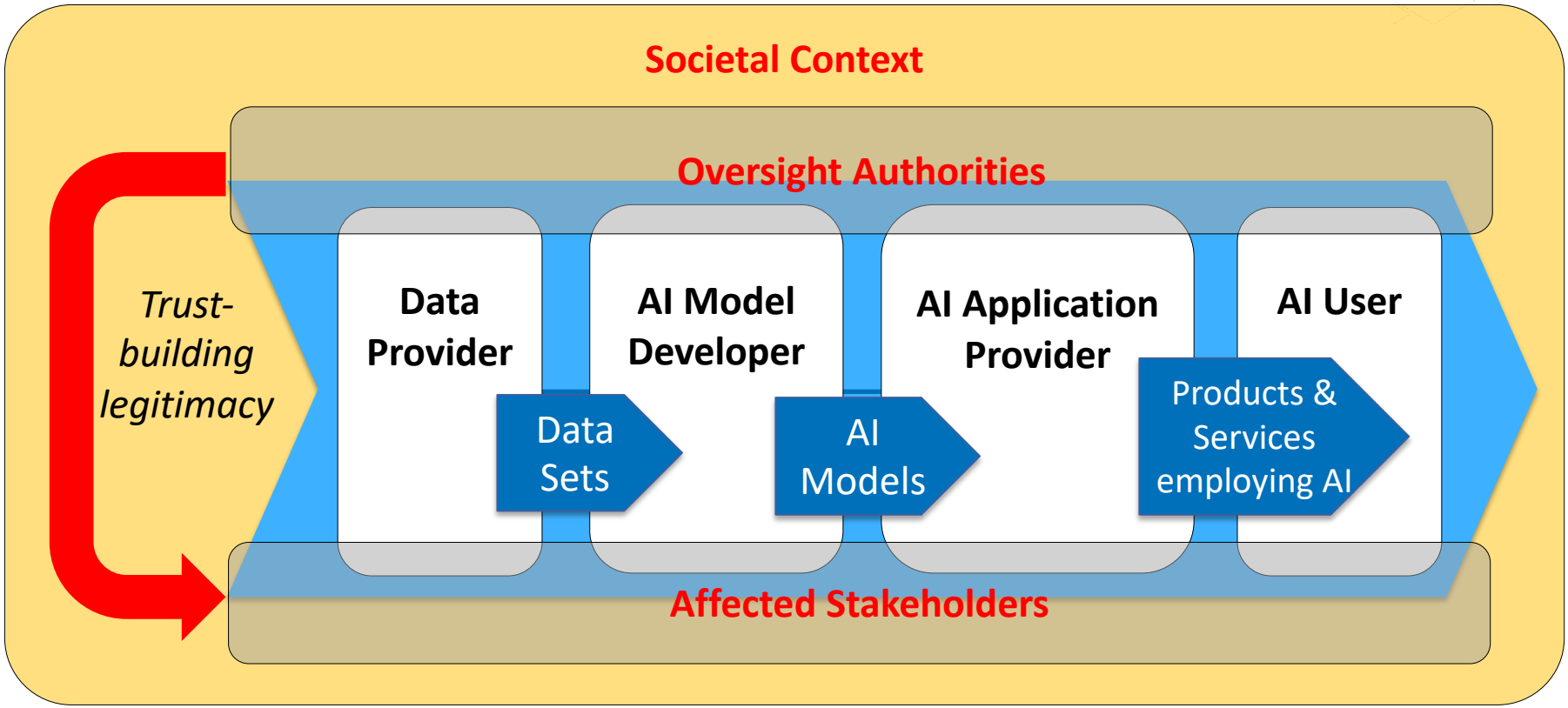
Ethical AI, alongside Lawful AI and Robust AI

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

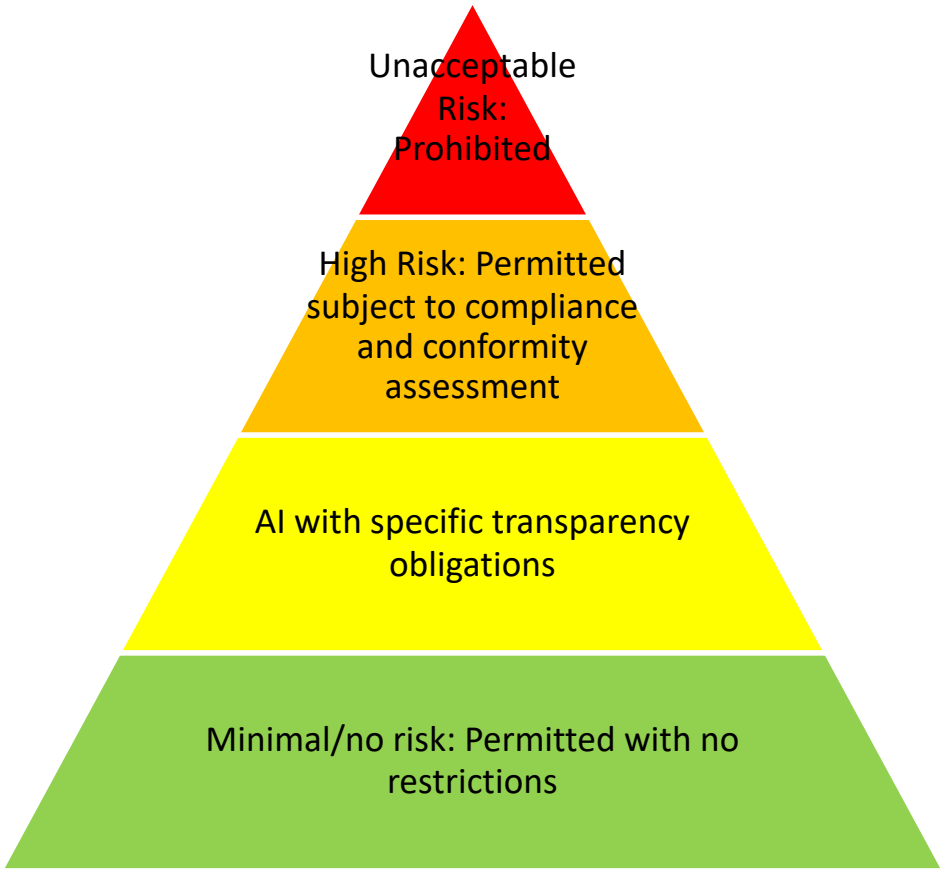
Principles to Practice: Governance in context



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?

Proposed Framework for EU AI Act



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

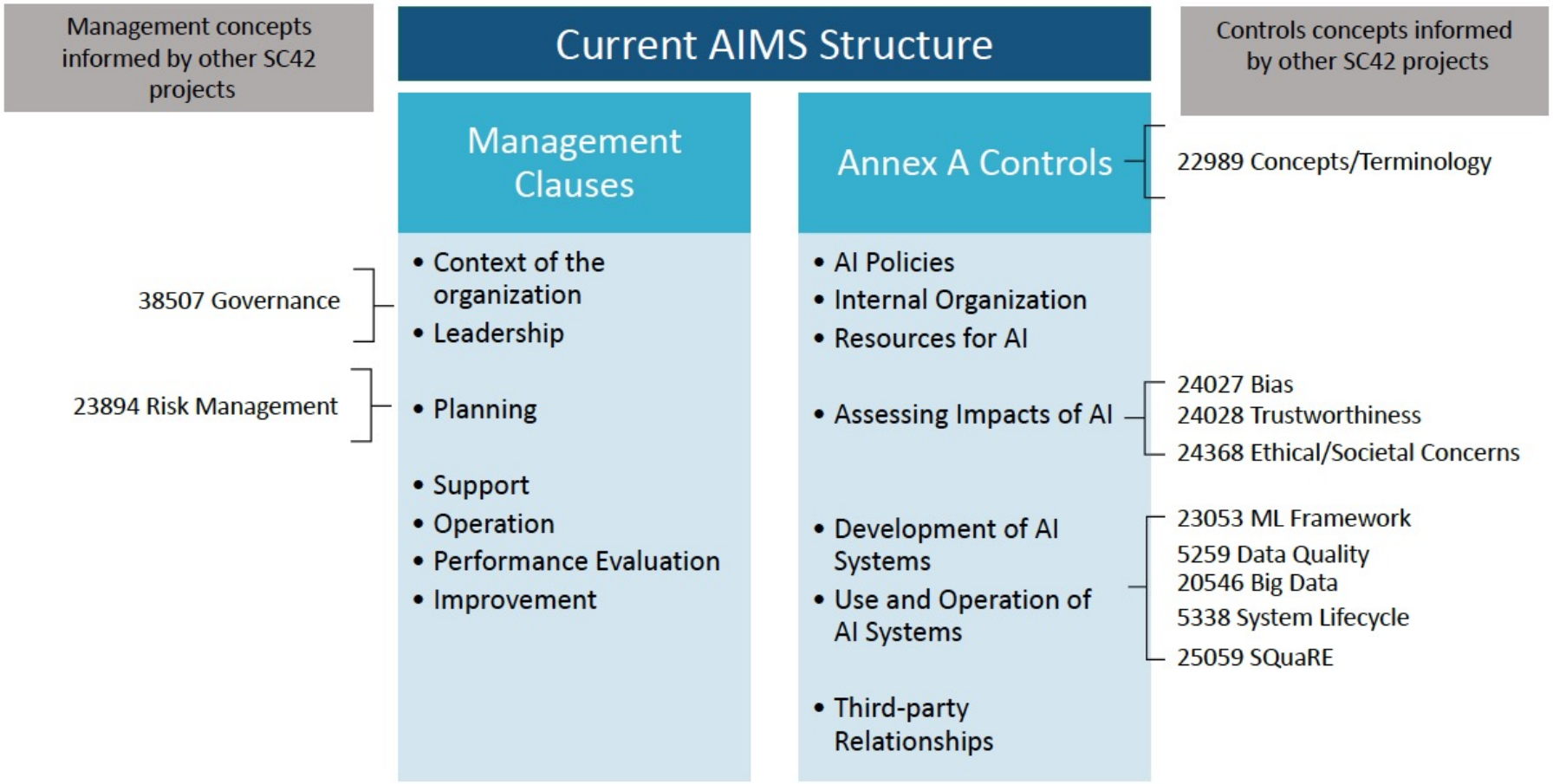
Provide legal certainty for innovation, promote public trust and support single market

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

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

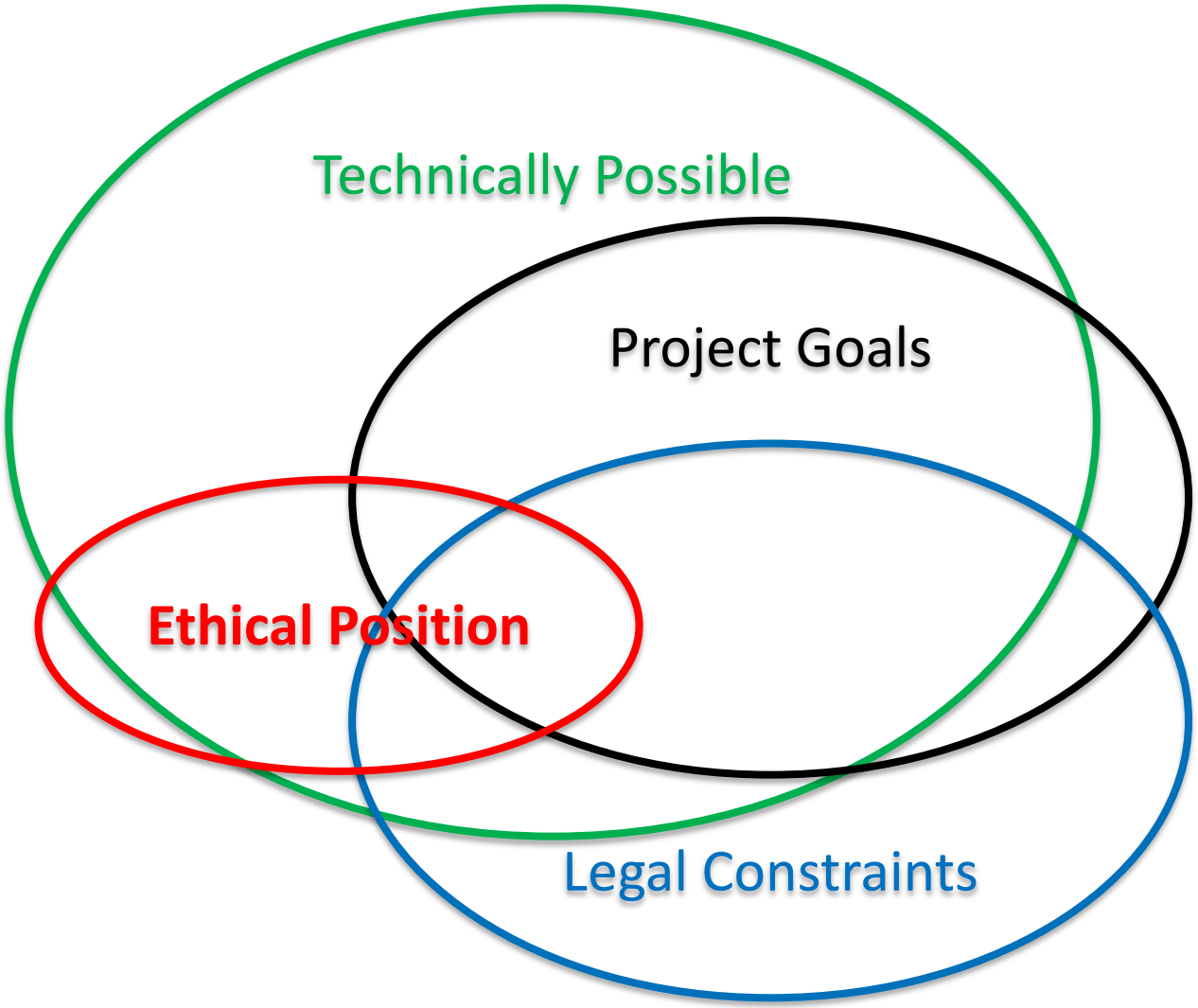


Considering Stakeholder in Addressing Tech Ethics

- **trustworthiness:** ability to meet stakeholder's expectations in a verifiable way
- **stakeholder:** any individual, group, or organization that can affect, be affected by, or perceive itself to be affected by a decision or activity [ISO/IEC 38500:2015]
- **Stakeholder types for Social Responsibility issues** [ISO 26000:2010]
    - Human Rights: everyone
    - Labour practices: workers
    - The environment: future generations
    - Fair operating practices: value chain customers and providers
    - Consumers
    - Local Community involvement and development

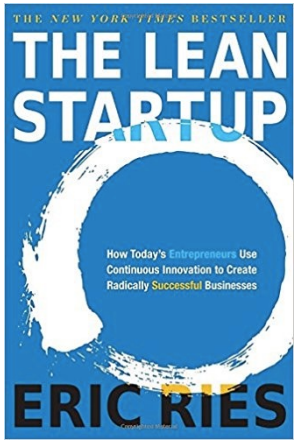
ISO JTC 1 IEC
INFORMATION TECHNOLOGY STANDARDS
SC 42 – Artificial Intelligence

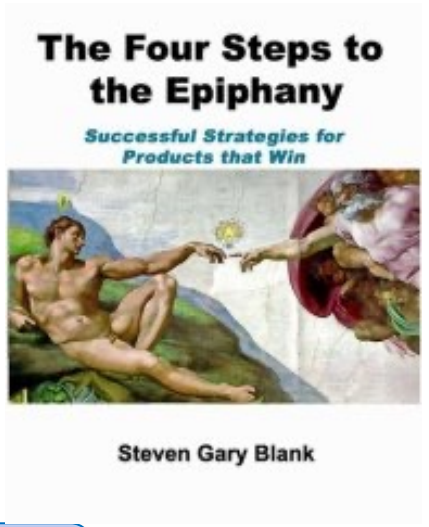
Ethics in a Technology Development Project



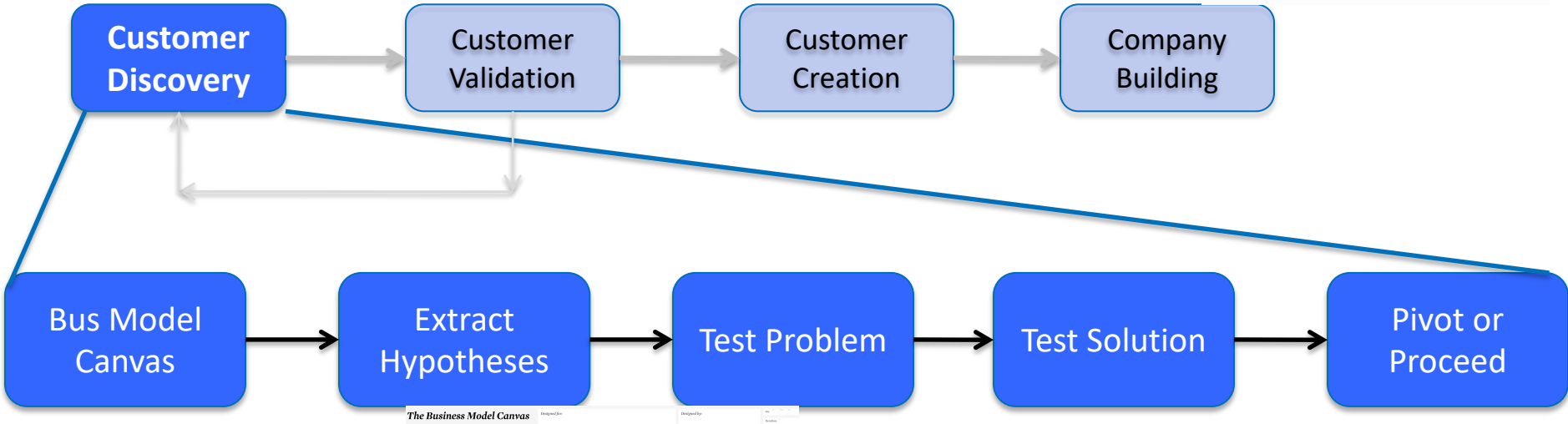
[IBM]

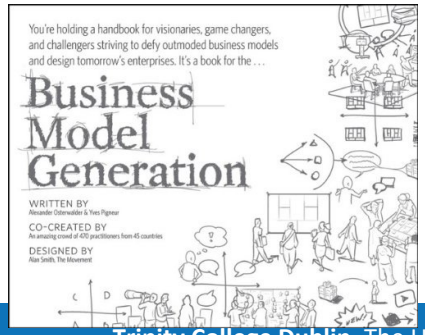
Data Hungry Innovation - "Silicon Valley" Methods

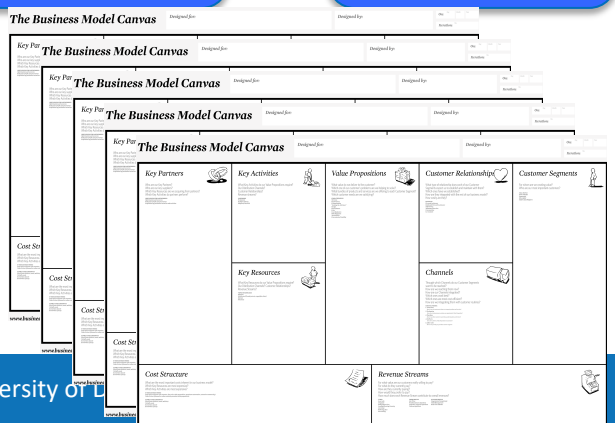




The Customer Development Process







How to make ethics part of the process?

Practicing Ethics in Responsible R&I

- Levels of practising ethics on responsible R&I (Brey, 2000):
  - *Disclosure*: exploration and identification of ethical impacts
  - *Theoretical*: frameworks to evaluate the impacts
  - *Application*: moral deliberation to overcome negative impacts

- *Disclosure* level neglected in current methodologies
- Need to:
  - Keep pace with **volume and speed** of innovation
  - **Accessible** to non-ethicist
    - R&I teams have an important perspective
    - R&I teams position to implement pivot to mitigate negative impact
  - Enabling a **collaborative** process

Ethics Canvas: Lightweight approach

- Ethic Canvas is a **methodology** for identifying, evaluating and resolving **ethical impacts** during **R&I** stages:
  - Formation of knowledge and concepts
  - Design of the technology
  - Prototyping and testing
  - Integration of R&I outcomes into society

- **Foster ethically informed technology design** by engaging R&I teams with the ethical impacts

- **Collaborative brainstorming tool** with two aims:
  - Help innovators identify, discuss and articulate possible ethical impacts
  - Bring about *pivots* in the design

https://ethicscanvas.org

Considerations on Ethical Impacts of Technology

- Changes in individual **behaviour**
- Relationships between **individuals**
- Relationships between **collective actors**
- **Relationships** *between* individuals and collectives
- Impact in the **public sphere, on worldviews**
- Impact of technology **failure**
- Impacts on the **environment** and production processes

Ethics Canvas

Project Title:

Date:

Individuals affected

Who use your product or service?
Who are affected by it's use?
Are they men/women, of different ages, etc.?

1

Behaviour

How might people's behaviour change because of your product or service?
Their habits, time-schedules, choice of activities, etc.?

3

Relations

How might relations between people and groups change because of your product or service? Between friends, family-members, co-workers, etc.?

4

What can we do?

What are the most important ethical impacts you found?
How can you address these by changing your design, organisation, or by proposing broader changes?

9

Worldviews

How might people's worldviews be affected by your product or service?
Their ideas about consumption, religion, work, etc.?

5

Group Conflicts

How might group conflict arise or be affected by your product or service?
Could it disciminate between people, put them out of work, etc.?

6

Groups affected

Which groups are involved in the design, production, distribution and use of your product or service?
Which groups might be affected by it?
Are these work-related organisation, interest groups, etc.?

2

Product or Service Failure

What are potential negative impact of your product or service failing to operate or to be used as intended?
What happens with technical errors, security failures, etc.?

7

Problematic Use of Resources

What are potential negative impacts of the consumption of resources relating to your project?
What happens with its use of energy, personal data, etc.?

8

Stage 1: Identify the Relevant Stakeholders

Who might be affected by application– be **inclusive**

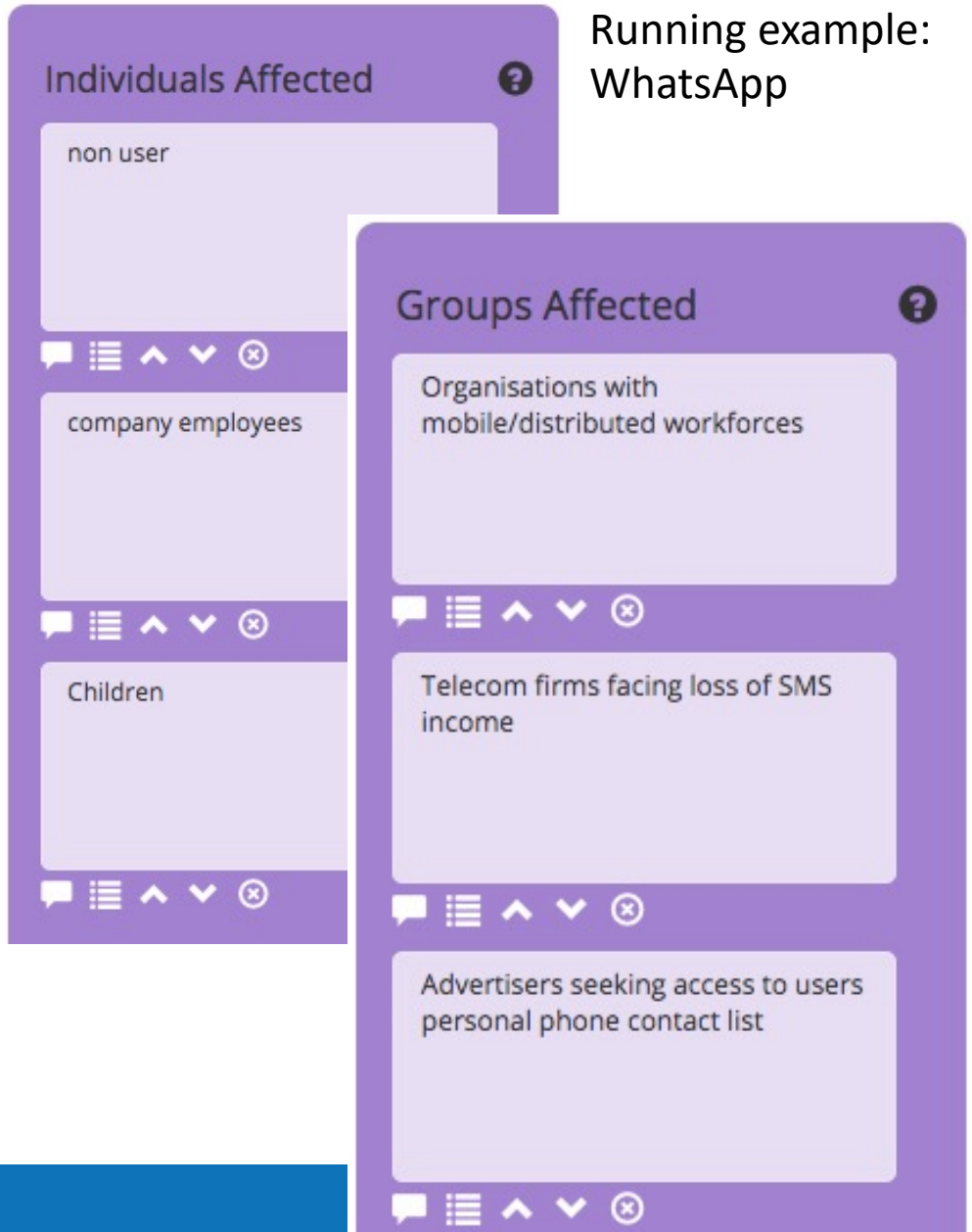
Individuals: Who use your product or service? Who are affected by it's use?

e.g are they of different genders, of different ages, etc.?

Groups: Which groups are involved in the design, production, distribution and use of your product or service?

Which groups might be affected by it?

e.g. are these work-related organisation, interest groups, etc.?



Running example: WhatsApp

Stage 2: Identifying Ethical Impacts

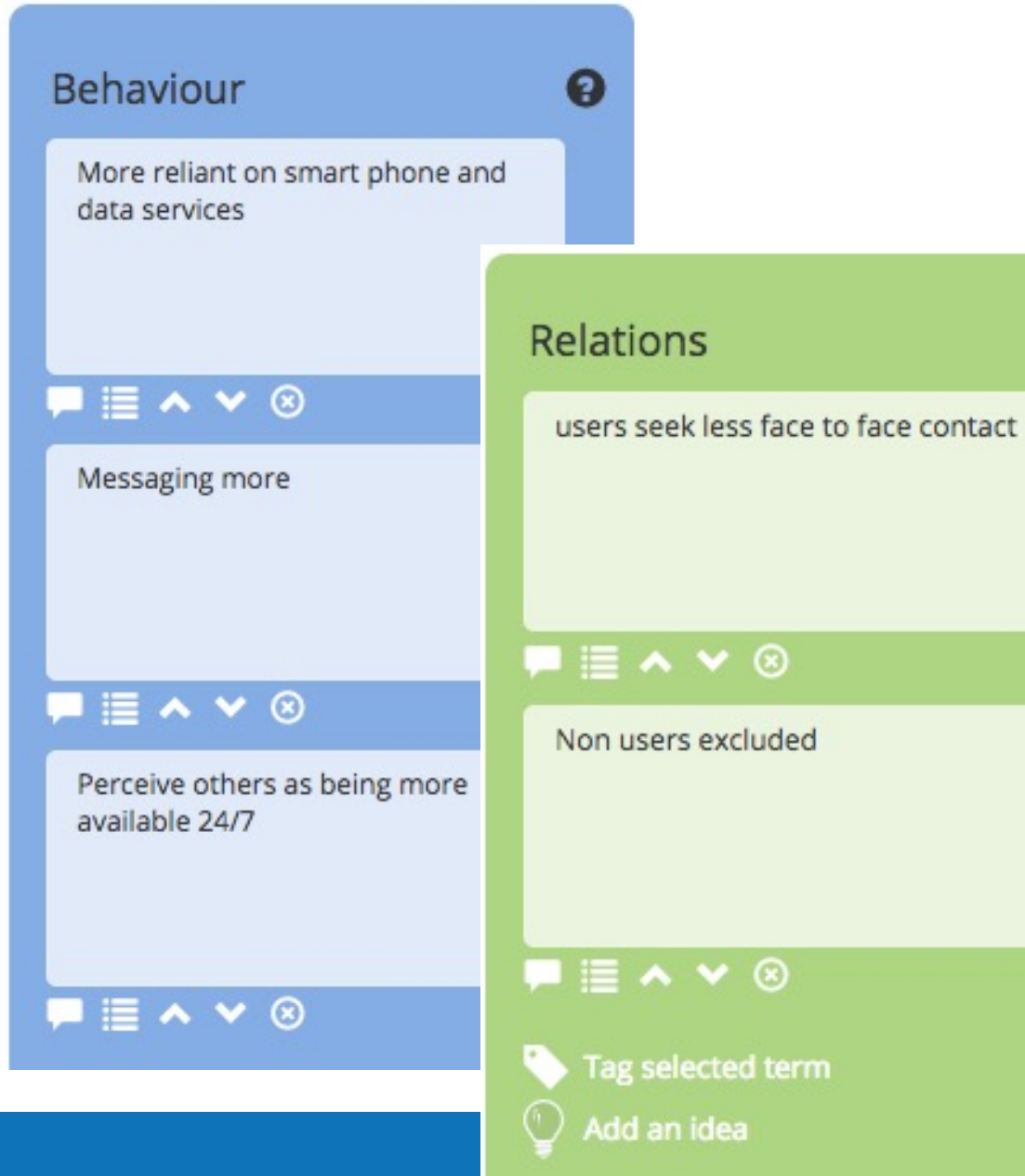
First, **'micro' impacts** are captured by the canvas, i.e. on everyday lives of people using and living with the application

Behaviour: How might people's behaviour change because of your product or service?

e.g. habits, time-schedules, choice of activities, etc.?

Relations: How might relations between people and groups change?

e.g. between friends, family members, co-workers, etc.?



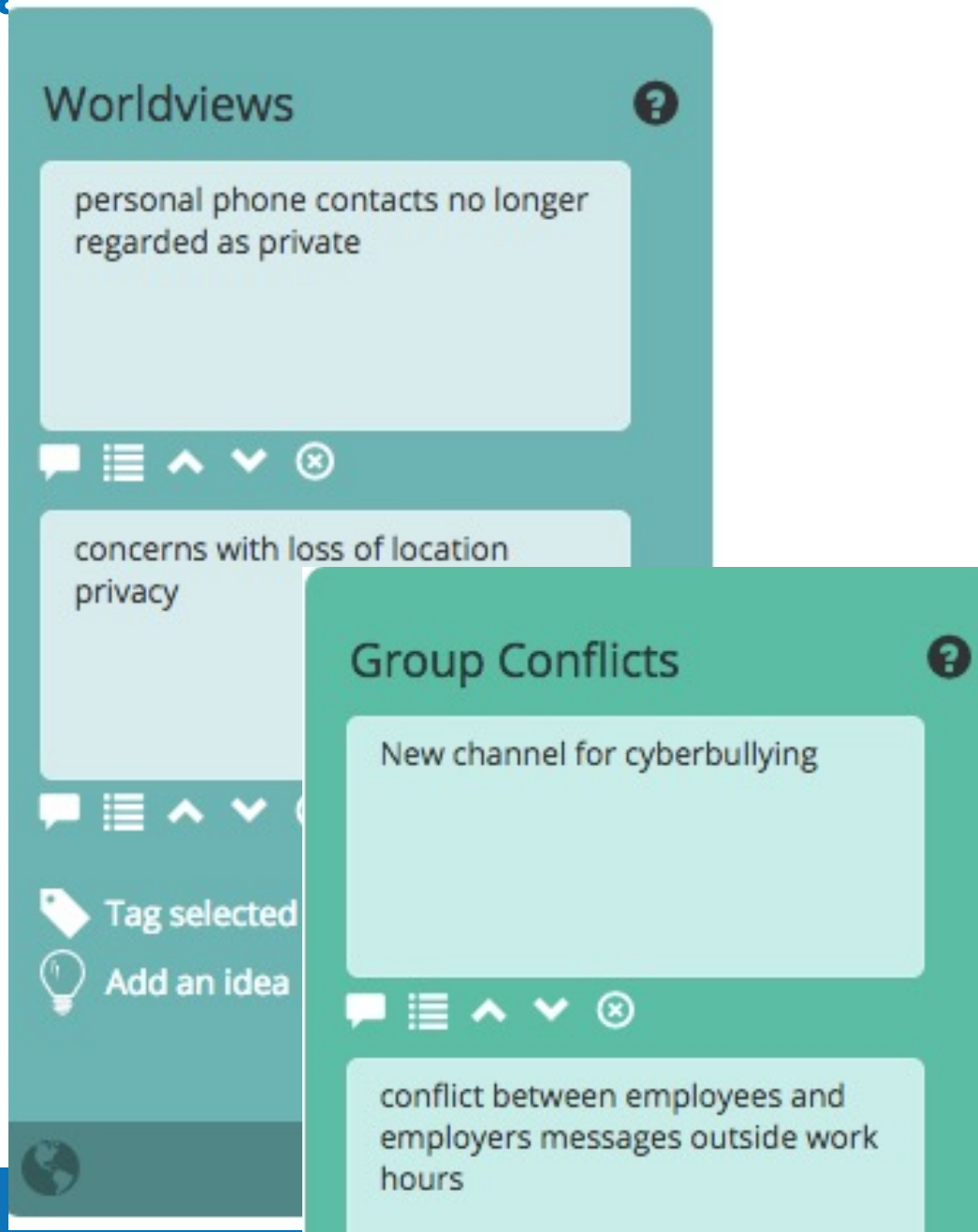
Stage 2: Identifying Ethical Impacts

Next **'macro' impacts** need to be considered.

These surpass individual's impacts - pertain *to* *collective*, social structures instead, e.g. related to political structures or cultural value-systems.

How might people's **Worldviews** be affected by your product or service? *e.g. their ideas about consumption, religion, work, etc.?*

Social conflicts: How might **Group Conflict** arise or be affected? *e.g.discriminate between people, put them out of work, etc.?*

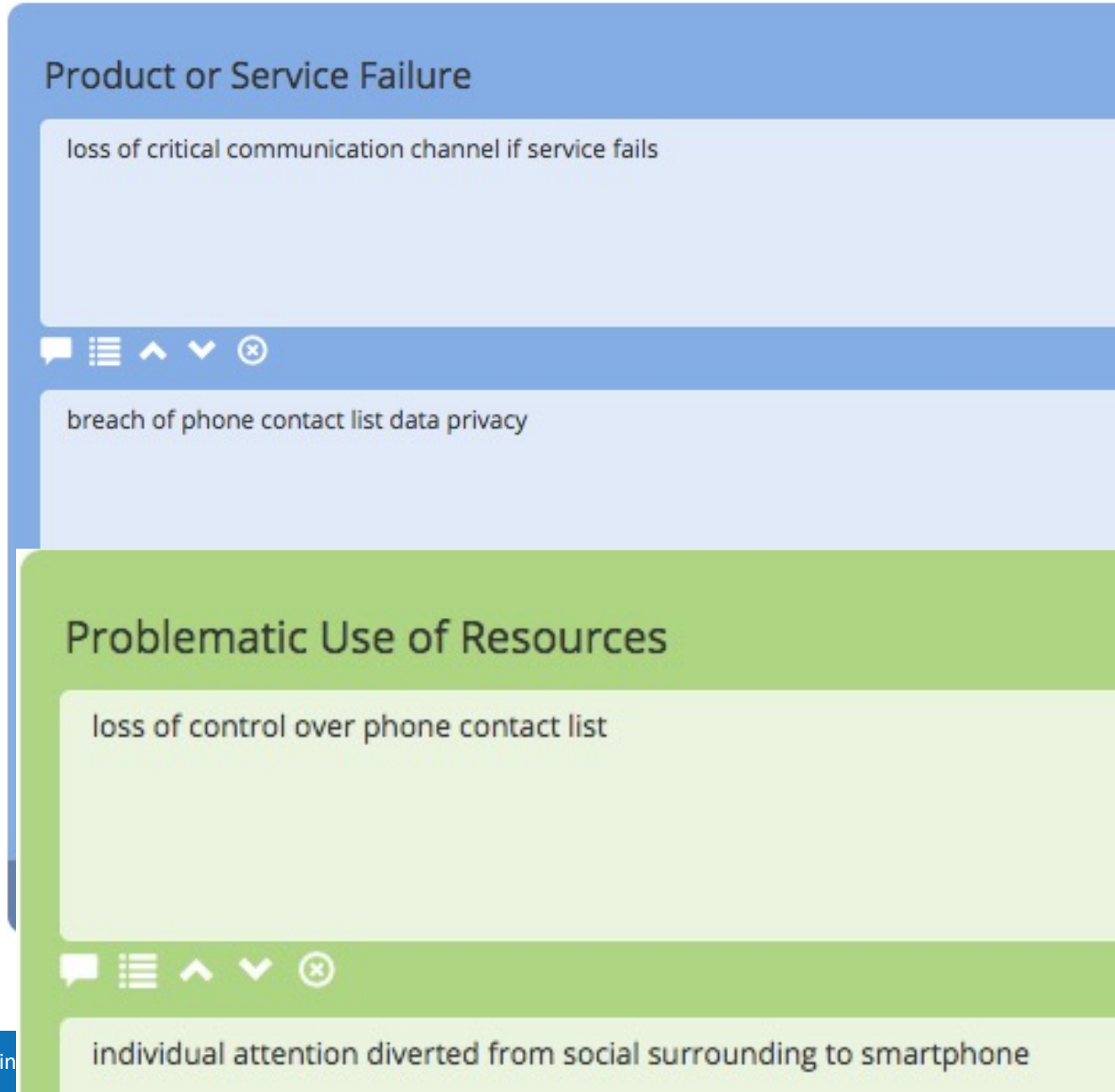


Stage 2: Identifying Ethical Impacts

Aspects that *indirectly* impact our lives..

Potential negative impact of your **product or service failure**? e.g. what happens with technical errors, security failures, etc.?

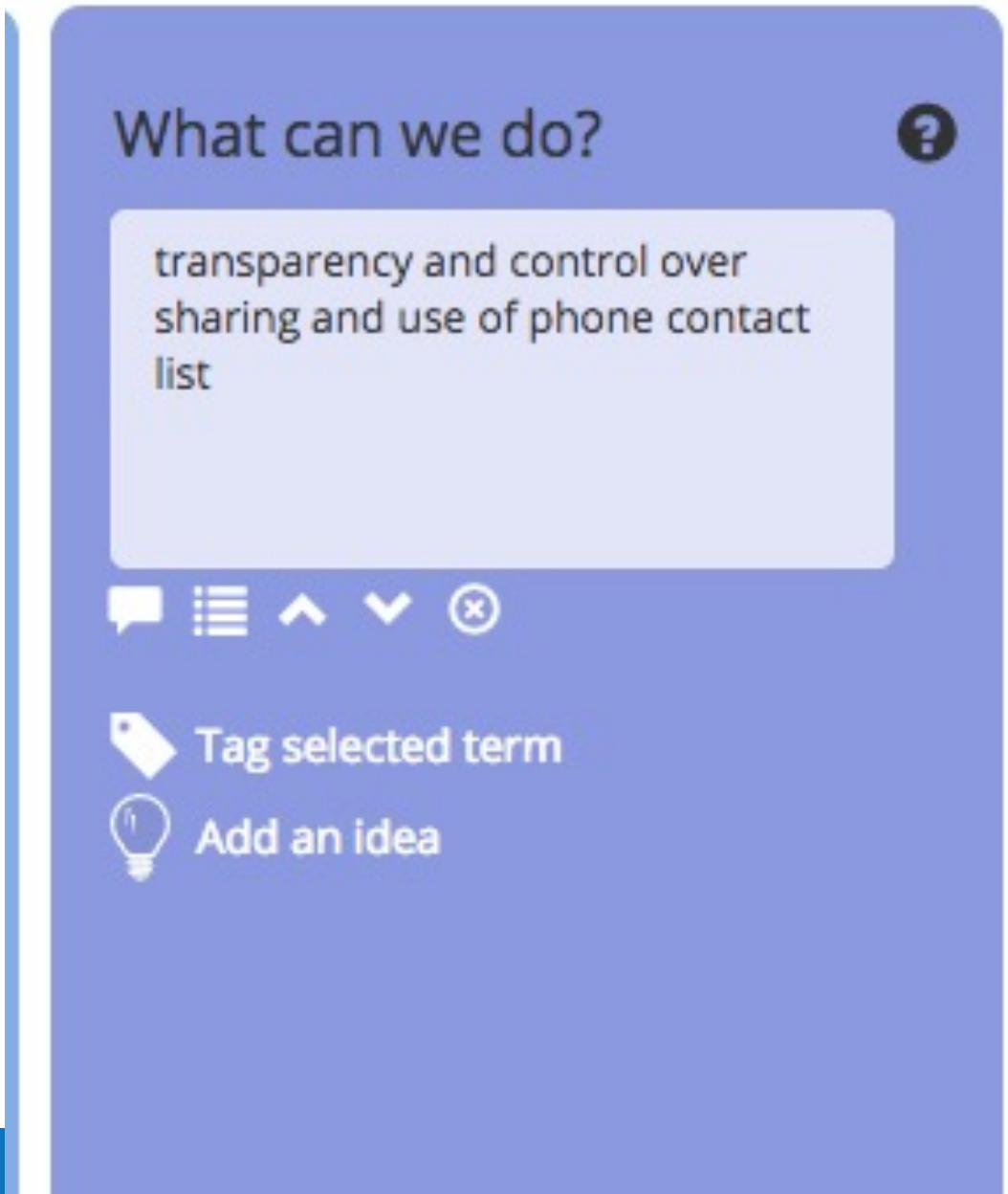
Potential negative impacts of the **consumption of resources** relating to your project? e.g. what happens with its use of energy, personal data, etc.?



Stage 3: How to Address Ethical Impacts

What are the most important ethical impacts you found?

How can you address these by pivoting your design, organisation, or by proposing broader changes?



Ethics Canvas, Group: Title: Date:

Individuals Affected:
-

Behaviour:
-

Relations:
-

What can we do?:
-

Worldviews:
-

Group Conflicts:
-

Groups affected:
-

Product or Service Failure:
-

Problematic Use of Resources:
-

Ethics Canvas, Group: Title: Microwave Example Date:

Individuals Affected:

Consumer of food

Behaviour:

Less time preparing meals

Easier to live singly/independently

More consumption of ready meals

Relations:

Less family interaction at meal times

What can we do?:

Find other reasons to eat together as a family

Microwave fresh rather than processing meals

Switch to air fryer

Worldviews:

- More individualistic outlooks
- Devaluing food preparation and cooking skills

Group Conflicts:

?

Groups affected:

Cooked food vendors – less business

Fresh food vendors: more value in pre-processed food as convenience attractive to consumers

Product or Service Failure:

Only way of warming food

Microwave unit leaks

Problematic Use of Resources:

More processed food and packaging

Ethics Canvas, Group: Title: YouTube Date:

Individuals Affected:

Everyone accessing youtube

Children

Content posters

Behaviour:

More screen time due to recommendations

Access to violent or disturbing content

Access to age inappropriate content

Open to false messages/information

Open for harmful body images

Relations:

Less consuming video as a group

Less consuming same video as social contacts, less common experience to share

What can we do?:

- Green energy for data centres and networks
- Screen time reporting and rationing
- Better screen on inappropriate content

Worldviews:

- Increase in belief in conspiracy theories
- increase in extremist and polarized views

Group Conflicts:

fakenews and distortion of facts impact civic and democratic processes

Employer harms on content moderators

Displacement of local news sources

Groups affected:

- News providers
- Advertisers
- Content providers
- youTubers
- Content moderators

Product or Service Failure:

- Loss of advertising opportunities
- Loss of video for promoting services or providing information, e.g. how-tos

Problematic Use of Resources:

- Data center power consumption

The Ethics Canvas

- Canvas current version: 1.8

- Web version:

https://ethicscanvas.org

- License: Creative Commons Attribution Non-Commercial 3.0 Unported

- User Manual available at:
- https://www.ethicscanvas.org/download/handbook.pdf

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

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

Thank You