# Epistemic Risk reduction:

# regulating online spaces

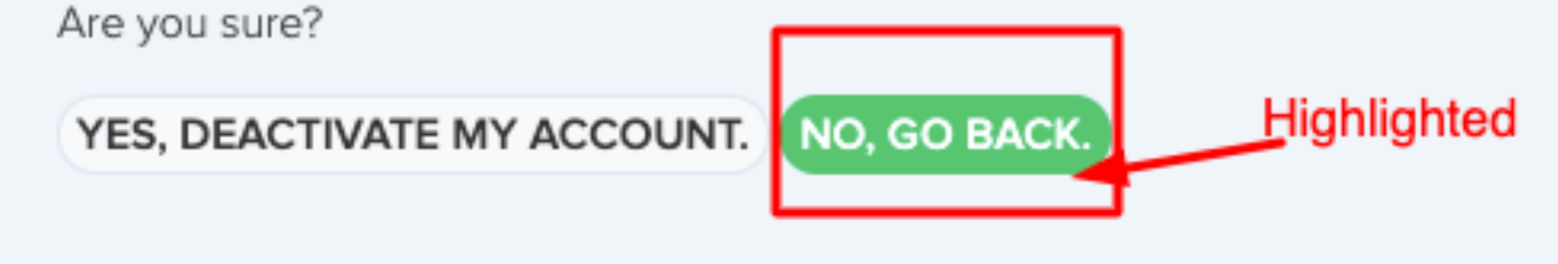
Carla Zoe Cremer, Human Information Processing Lab, Centre for the Governance of AI

Centre for the Governance of AI

# A framework

Opinion formation as behavior

Every environment promotes and curbs certain behaviors


Are you sure?

Carla Cremer, talk @ UK DCMS Committee Online Harms and Disinformation , July 2022

# Behavior in space

Spaces embed values and determine political outcomes.

Mark Beissinger (Princeton University Press, 2022).

# Epistemic Risk & Resilience

A longterm, tail-risk approach

Episteme – knowledge / wisdom

@noemidrexler

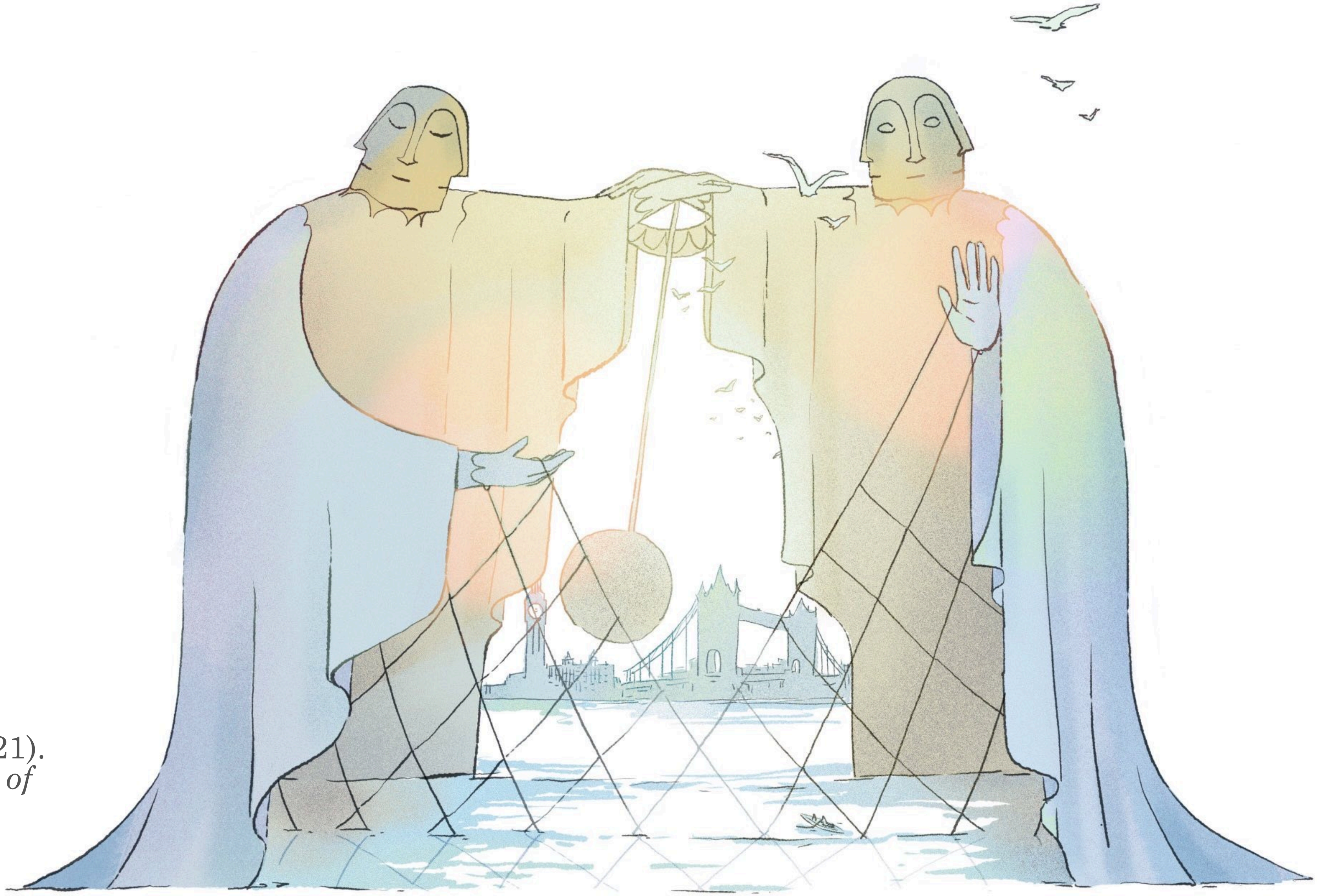
Carla Cremer, talk @ UK DCMS Committee Online Harms and Disinformation , July 2022

# Epistemic Risk

# &

# Resilience

Cremer, C. Z., & Kemp, L. (2021). *Democratising Risk: In Search of a Methodology to Study Existential Risk*



@magdalenadomeit

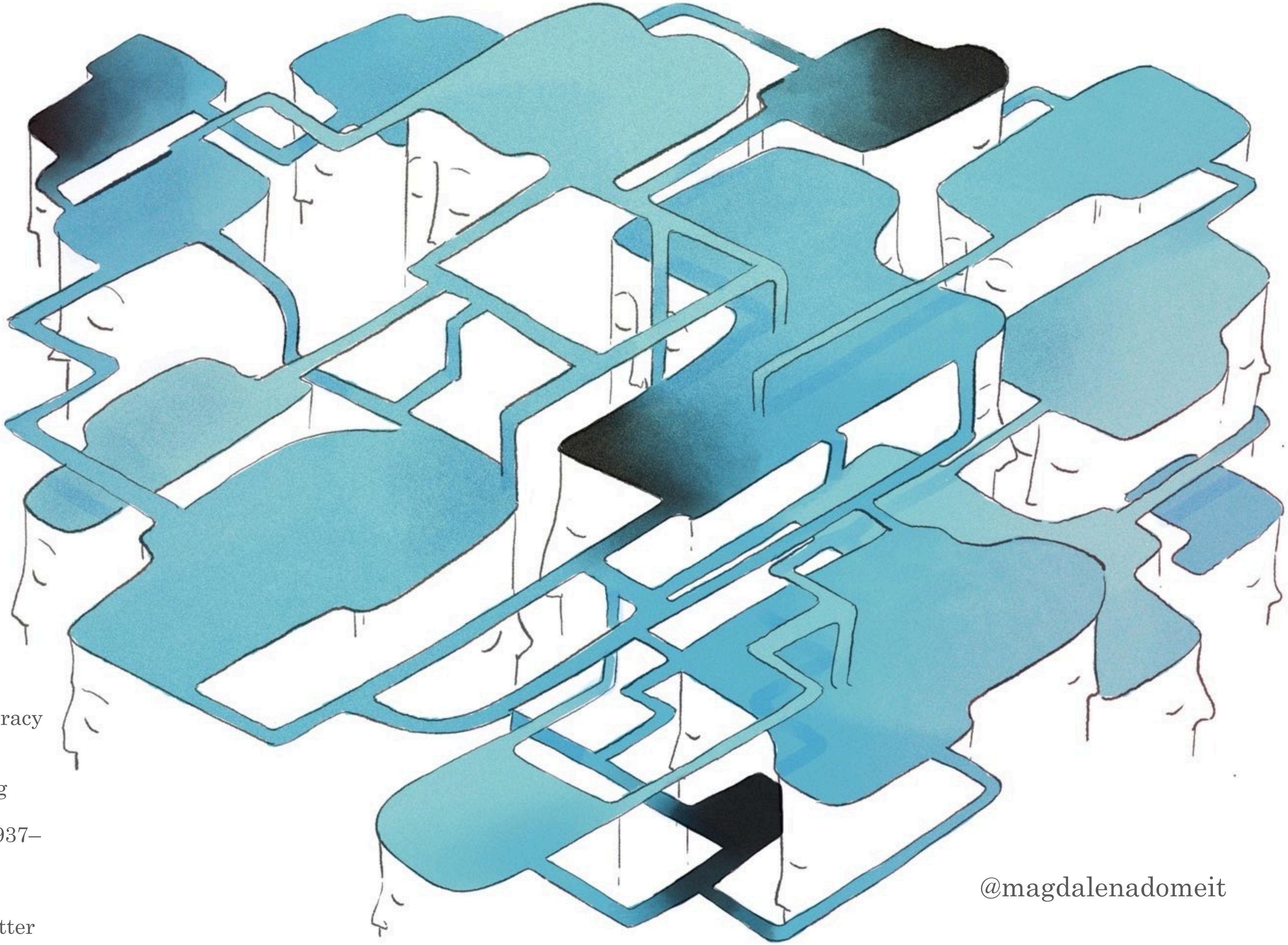
- Architectural features > Content
- Content-neutral interventions
- Psychology Research

@noemidrexler

# Conspiracy

# &

# Democracy



@magdalenadomeit

Schaeffer et al. July 2022; https://doi.org/10.1073/pnas.2203149119

Radnitz, S. (2022). Why Democracy Fuels Conspiracy Theories. *Journal of Democracy*, *33*(2), 147–161.

Cusimano, C., & Lombrozo, T. (2021). Reconciling scientific and commonsense values to improve reasoning. *Trends in Cognitive Sciences*, *25*(11), 937–949.

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50.

# Hard choices

and who gets to decide

Trade-offs:

- eg. harmful content vs self-curated feeds

Arbiters of politics:

- Scheck, Jeff Horwitz and Justin. 2021.
- *Querdenker* as test-bed

Productive polarization?

- Stray, 2022

*"This could be a good case study to inform how we tackle these problems in the future"*

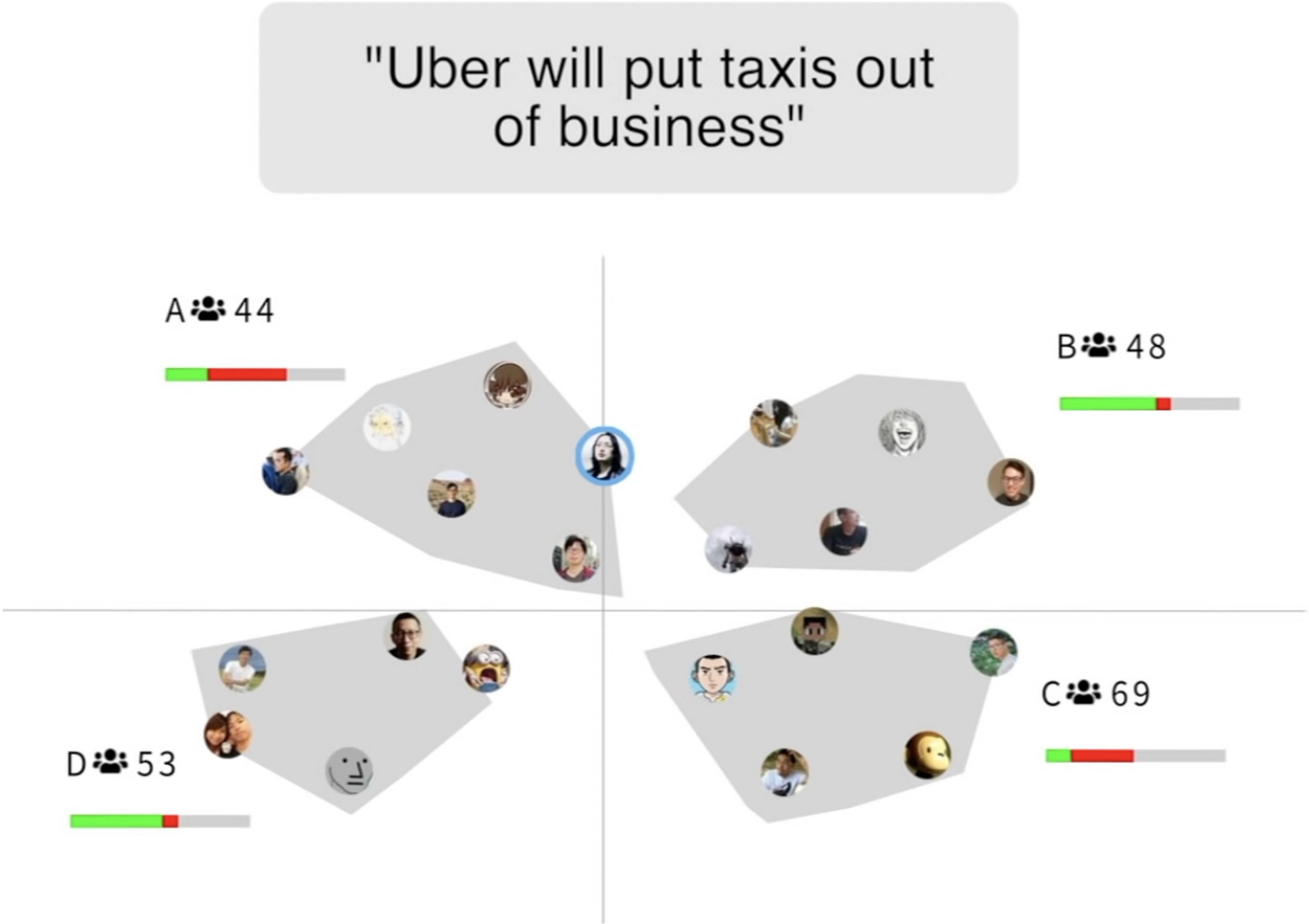
*"An individual can question election results. But when it's amplified by a movement, it can damage democracy. There is harm in the way movements shift norms and an understanding of collective truth."*

# Participation

is necessary

works

must be researched



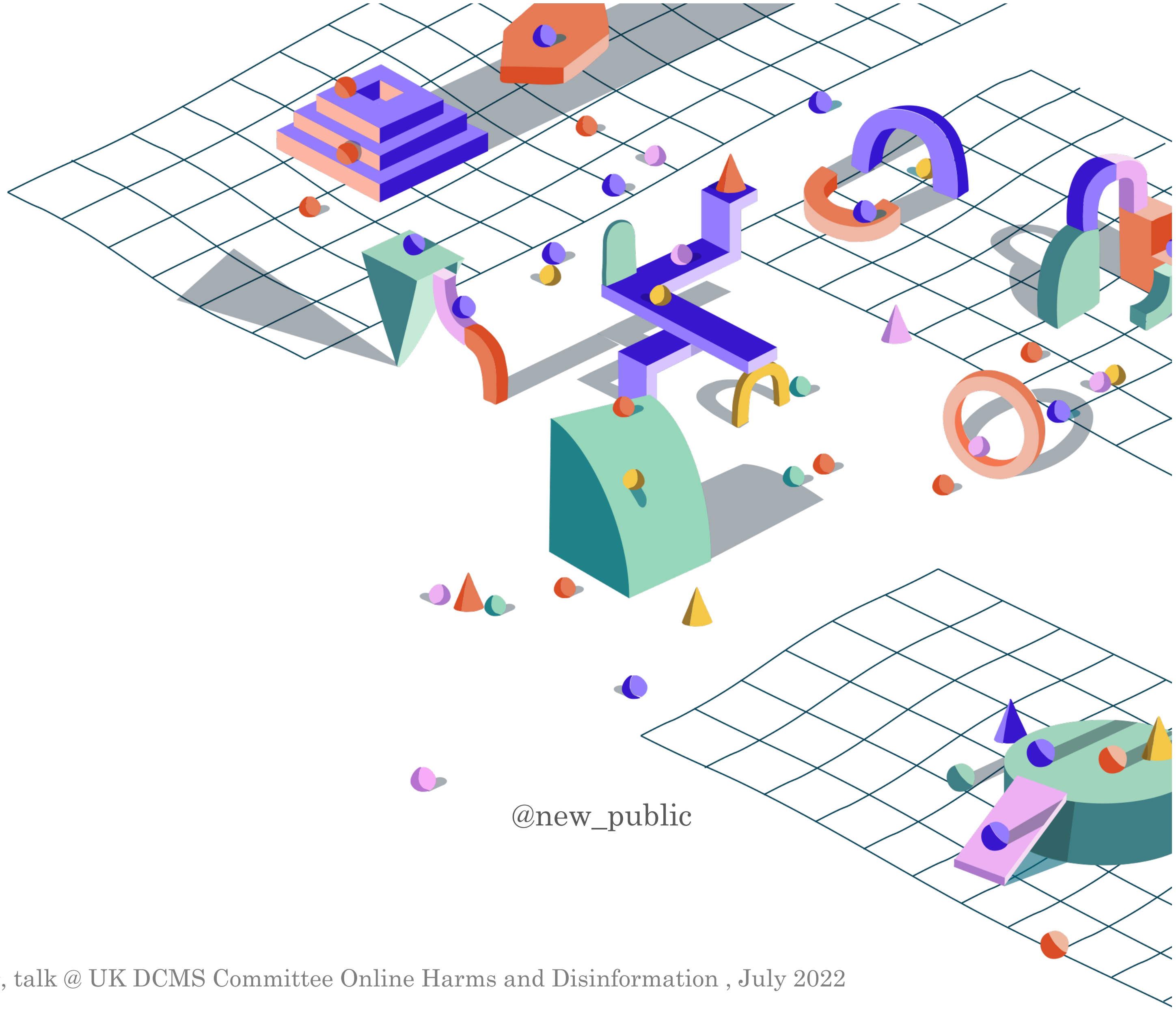
- Landemore (2017); vTaiwan, Landemore (2021), OECD (2020); Lee et al. (2019)

# In dire need of psychological theory

- Features facilitating cognitive frames;
  - Silva, A., et al. (2022). *BIT Report*, 36.
  - *New Public* Signals Research Overview
- What is democratically important content?
- Polarisation tipping points?
  - Stray, J. (2022)
- Algorithmic consensus finding
  - Koster, R et al. (2022). Human-centred mechanism design with Democratic AI. *Nature Human Behaviour*, 1–10.
  - *Pol.is & vTaiwan*; MIT Technology Review 2018

- Deliberation
- Benchmarks --- verification
- Markers --- correlates of ‘healthy’ conflict / discourse
- Metrics --- recommendation performance

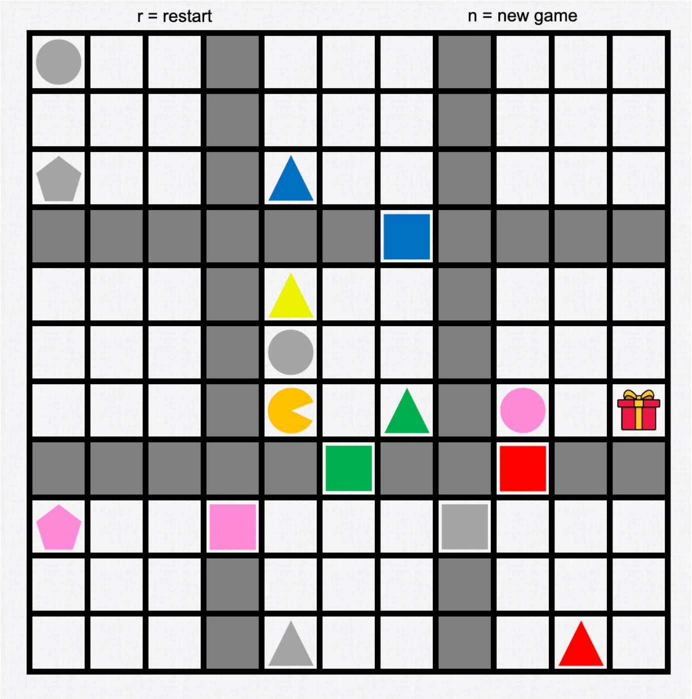
# We need a cognitive theory of online spaces



@new_public

# Human information processing lab

- truth-seeking games
- curriculum and information sharing
- validity must be tested in the wild

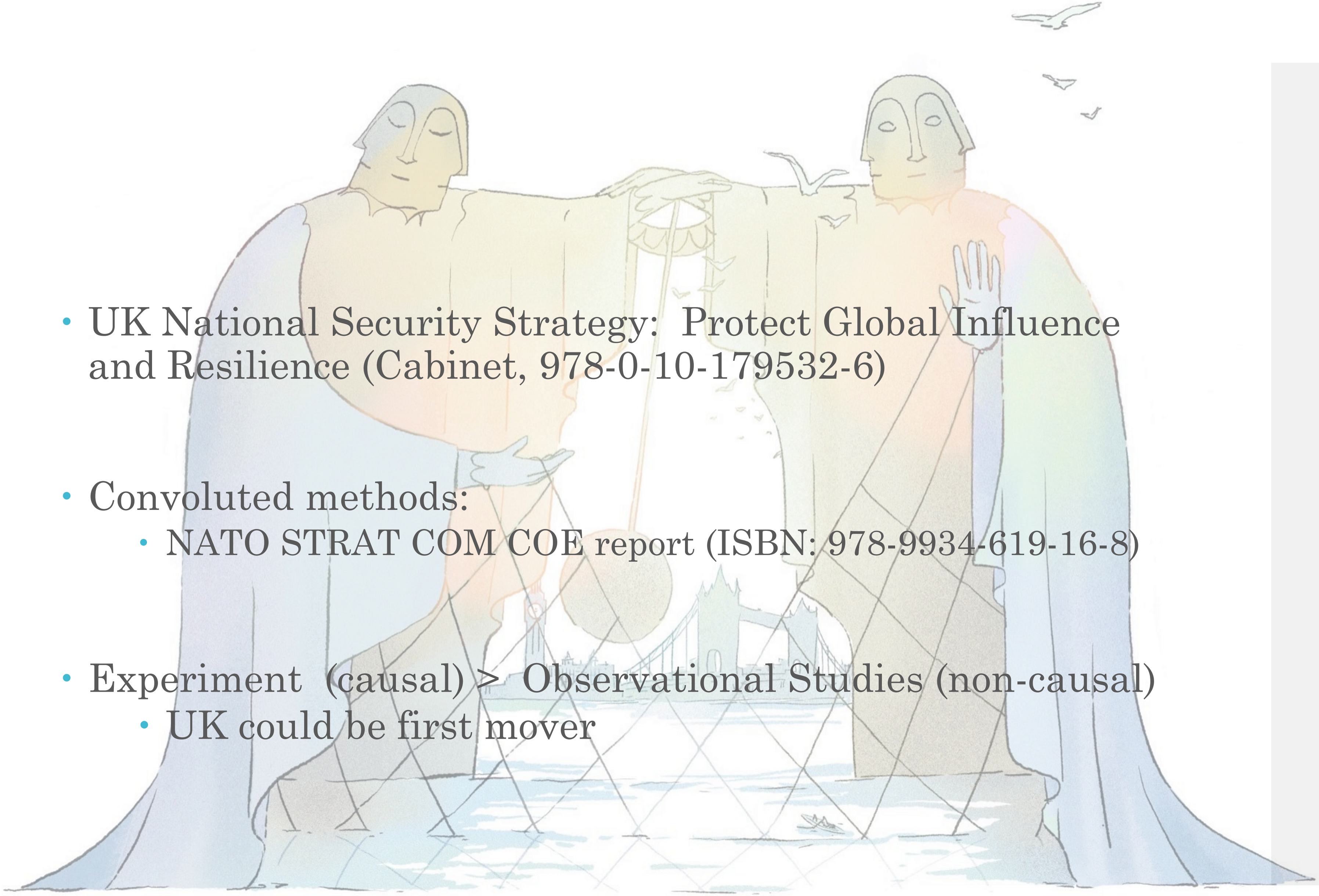


Dumbalska, 2022

# Insight,

# Security &

# Control



- UK National Security Strategy:  Protect Global Influence and Resilience (Cabinet, 978-0-10-179532-6)

- Convoluted methods:
  - NATO STRAT COM COE report (ISBN: 978-9934-619-16-8)

- Experiment  (causal) >  Observational Studies (non-causal)
  - UK could be first mover

# Suggestions

Epistemic Risk Reduction

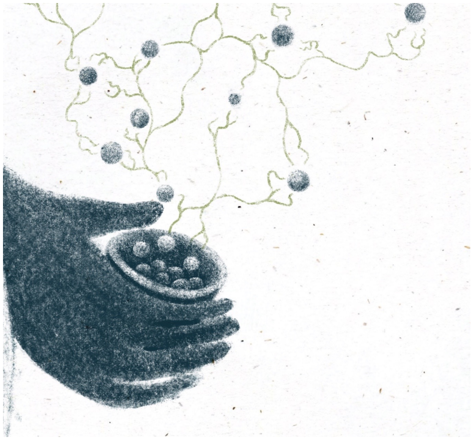
- Legal Requirement for research: effects of novel architectural features
- Compensations: users surveys and deliberations
- Participatory Architecture Assessments
  - OECD. 2020. *Innovative Citizen Participation and New Democratic Institutions* (March 2, 2021).
  - Cremer, C. Z., & Whittlestone, J. (2021). *Artificial Canaries*
  - Lee, M. et. al.(2019). *WeBuildAI: Participatory Framework for Algorithmic Governance.* Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1–35.

# Proactive technology assessments

Online Safety Bill

- Verification: assessing efforts relies on knowing what's possible

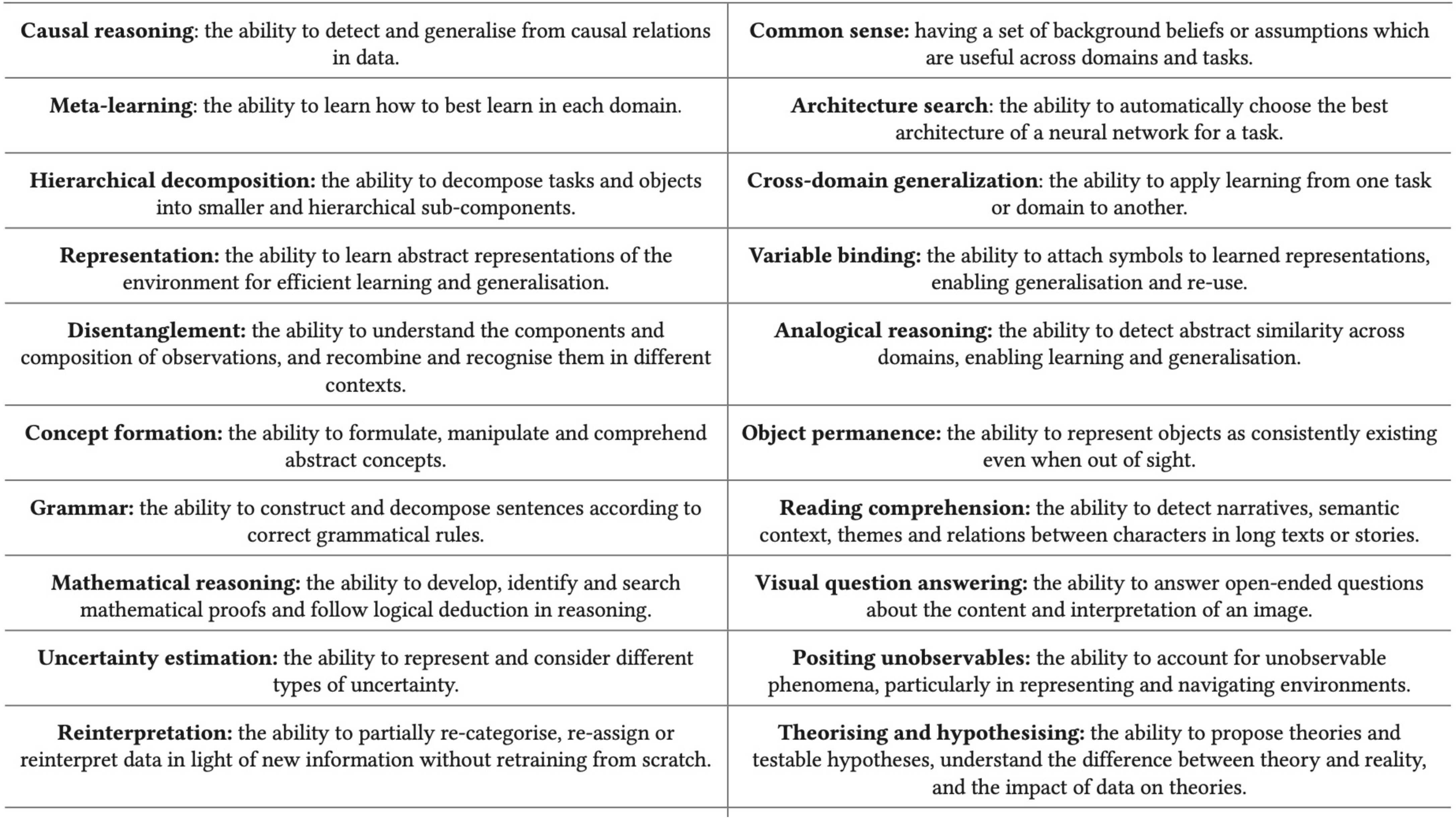
- Proactive technology: assessments
  - Expert identification
  - Limitation assessment -- > identify misapplication



Cremer, C. Z. (2021). Deep Limitations? Examining Expert Disagreement over Deep Learning. *Progress in Artificial Intelligence, Springer*

Cremer, C. Z., & Whittlestone, J. (2021). Artificial Canaries: Early Warning Signs for Anticipatory and Democratic Governance of AI

TABLE I. LIMITATIONS OF DEEP LEARNING AS PERCEIVED AND NAMED BY EXPERTS FOUND IN [11]



| | |
|---|---|
| **Causal reasoning**: the ability to detect and generalise from causal relations in data. | **Common sense:** having a set of background beliefs or assumptions which are useful across domains and tasks. |
| **Meta-learning**: the ability to learn how to best learn in each domain. | **Architecture search**: the ability to automatically choose the best architecture of a neural network for a task. |
| **Hierarchical decomposition:** the ability to decompose tasks and objects into smaller and hierarchical sub-components. | **Cross-domain generalization**: the ability to apply learning from one task or domain to another. |
| **Representation:** the ability to learn abstract representations of the environment for efficient learning and generalisation. | **Variable binding:** the ability to attach symbols to learned representations, enabling generalisation and re-use. |
| **Disentanglement:** the ability to understand the components and composition of observations, and recombine and recognise them in different contexts. | **Analogical reasoning:** the ability to detect abstract similarity across domains, enabling learning and generalisation. |
| **Concept formation:** the ability to formulate, manipulate and comprehend abstract concepts. | **Object permanence:** the ability to represent objects as consistently existing even when out of sight. |
| **Grammar:** the ability to construct and decompose sentences according to correct grammatical rules. | **Reading comprehension:** the ability to detect narratives, semantic context, themes and relations between characters in long texts or stories. |
| **Mathematical reasoning:** the ability to develop, identify and search mathematical proofs and follow logical deduction in reasoning. | **Visual question answering:** the ability to answer open-ended questions about the content and interpretation of an image. |
| **Uncertainty estimation:** the ability to represent and consider different types of uncertainty. | **Positing unobservables:** the ability to account for unobservable phenomena, particularly in representing and navigating environments. |
| **Reinterpretation:** the ability to partially re-categorise, re-assign or reinterpret data in light of new information without retraining from scratch. | **Theorising and hypothesising:** the ability to propose theories and testable hypotheses, understand the difference between theory and reality, and the impact of data on theories. |

- Cremer, C. Z. (2021). Deep Limitations? Examining Expert Disagreement over Deep Learning. *Progress in Artificial Intelligence, Springer*.

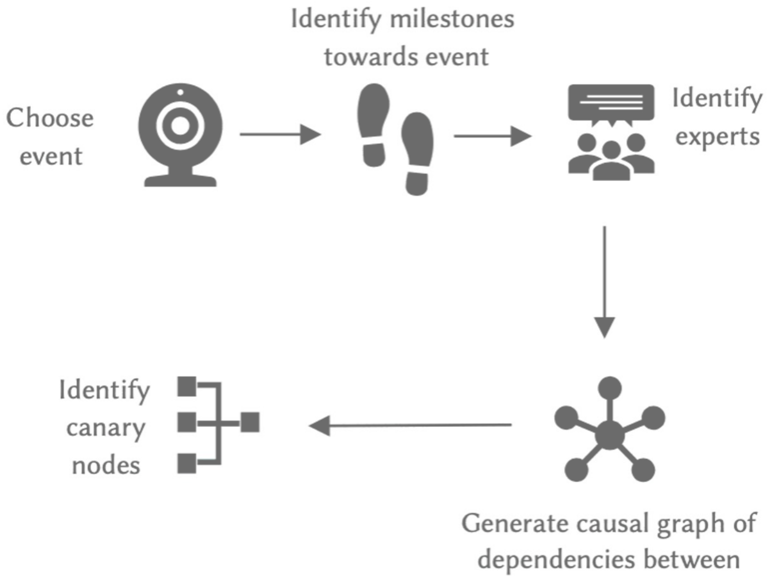
# Preemptive technology assessments

Expert forecasting

(could be enforced by regulator)

## Assessing technological advance:

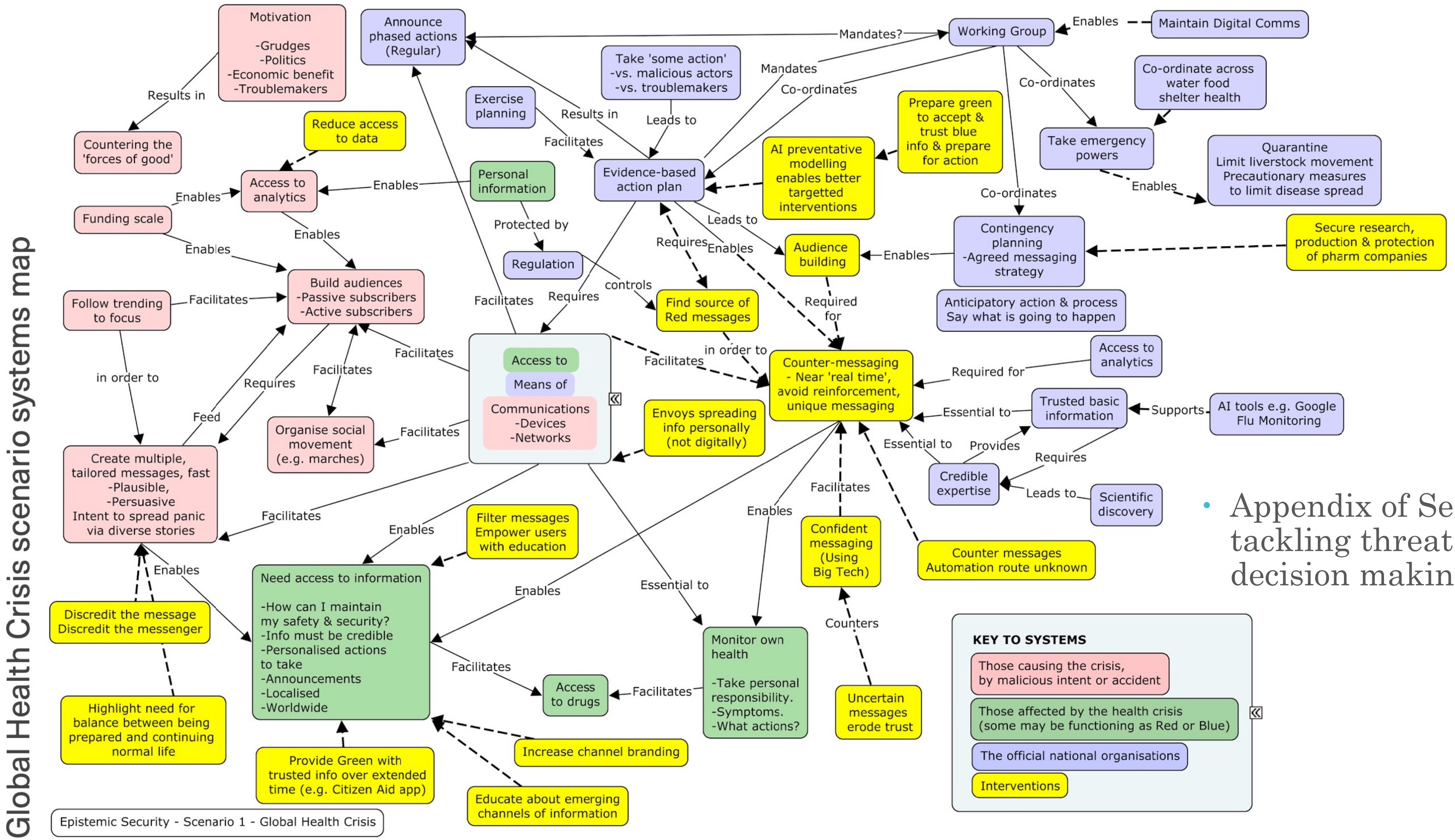
- Cremer, C. Z., & Whittlestone, J. (2021).



Choose event → Identify milestones towards event → Identify experts → Generate causal graph of dependencies between → Identify canary nodes

## Assessing impact of technological advance:

- red teaming, pre-mortems, second-order effects
- Seger, E., Avin, S., Pearson, G., Briers, M., O Heigeartaigh, S., & Bacon, H. (2020).



Global Health Crisis scenario systems map

Motivation

-Grudges
-Politics
-Economic benefit
-Troublemakers

Results in

Countering the 'forces of good'

Announce phased actions (Regular)

Mandates?

Working Group

Enables

Maintain Digital Comms

Take 'some action' -vs. malicious actors -vs. troublemakers

Mandates

Co-ordinates

Co-ordinates

Co-ordinate across water food shelter health

Exercise planning

Results in

Leads to

Facilitates

Reduce access to data

Access to analytics

Enables

Personal information

Evidence-based action plan

AI preventative modelling enables better targetted interventions

Prepare green to accept & trust blue info & prepare for action

Take emergency powers

Quarantine
Limit liverstock movement
Precautionary measures
to limit disease spread

Enables

Enables

Funding scale

Enables

Enables

Protected by

Regulation

Leads to

Requires

Enables

Audience building

Enables

Co-ordinates

Contingency planning -Agreed messaging strategy

Secure research, production & protection of pharm companies

Build audiences
-Passive subscribers
-Active subscribers

Follow trending to focus

Facilitates

controls

Facilitates

Requires

Find source of Red messages

Required for

Anticipatory action & process
Say what is going to happen

Facilitates

Facilitates

Access to

Means of

Communications
-Devices
-Networks

in order to

Facilitates

Counter-messaging
- Near 'real time',
avoid reinforcement,
unique messaging

Required for

Access to analytics

in order to

Requires

Feed

Organise social movement (e.g. marches)

Facilitates

Envoys spreading info personally (not digitally)

Essential to

Trusted basic information

Supports

AI tools e.g. Google Flu Monitoring

Essential to

Provides

Requires

Create multiple, tailored messages, fast
-Plausible,
-Persuasive
Intent to spread panic via diverse stories

Credible expertise

Leads to

Scientific discovery

Facilitates

Facilitates

Enables

Filter messages
Empower users
with education

Enables

Confident messaging (Using Big Tech)

Counter messages
Automation route unknown

Enables

Need access to information

-How can I maintain my safety & security?
-Info must be credible
-Personalised actions to take
-Announcements
-Localised
-Worldwide

Enables

Essential to

Discredit the message
Discredit the messenger

Counters

Monitor own health

-Take personal responsibility.
-Symptoms.
-What actions?

Facilitates

Access to drugs

Facilitates

Uncertain messages erode trust

Highlight need for balance between being prepared and continuing normal life

Increase channel branding

Provide Green with trusted info over extended time (e.g. Citizen Aid app)

Educate about emerging channels of information

KEY TO SYSTEMS

Those causing the crisis, by malicious intent or accident

Those affected by the health crisis (some may be functioning as Red or Blue)

The official national organisations

Interventions

Epistemic Security - Scenario 1 - Global Health Crisis

- Appendix of Seger et al 2020.: tackling threats to informed decision making

# Suggestions

Online Safety Bill

- Reporting: require reports on significant architectural changes
- Research: require research collaborations on architectural choices
- Research: require *experimental* access
- Proactive technology assessments

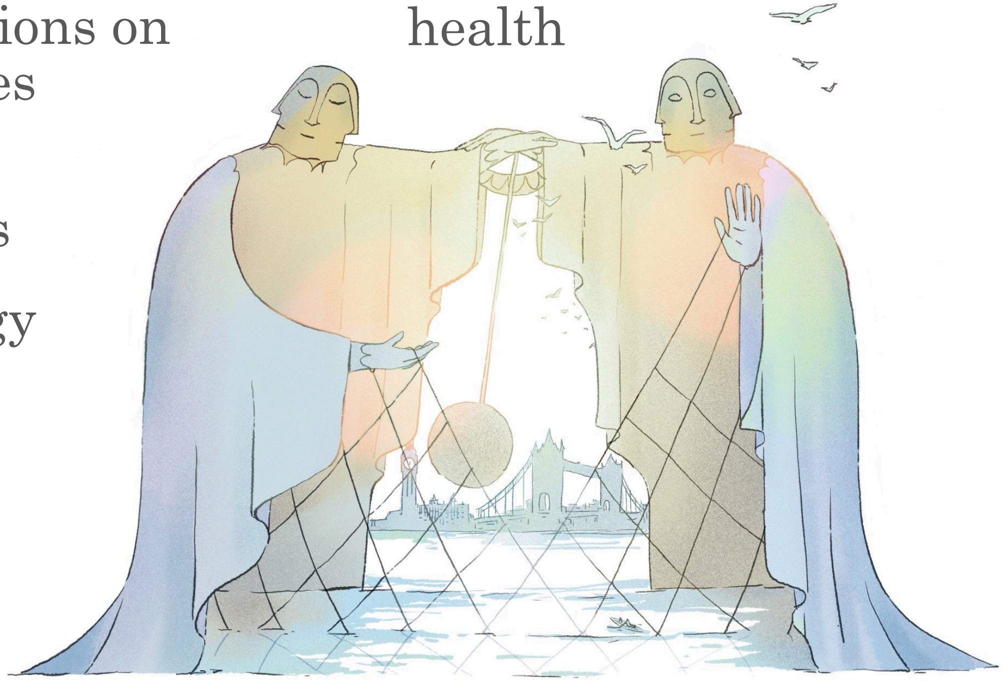
# Suggestions

Epistemic Risk Reduction

- Foster data-altruism
- Foster and train open-source journalism / task force
  - OSINT / bellincat
  - sudan media capacity building project
  - Wikimedians, Wiki Education Foundation
  - reward mechanisms

# Summary

- Reporting: require reports on significant architectural changes
- Research: require research collaborations on architectural choices
- Research: require *experimental* access
- Proactive technology assessments
- Participatory Architecture Assessments
- Cognition research: markers of epistemic health



@magdalenadomeit

carla.cremer@queens.ox.ac.uk ;

@CarlaZoeC

Papers:

https://carlacremer.github.io/research/

Illustrations:

https://magdalenaadomeit.com/

https://www.governance.ai/team/noemi-dreksler