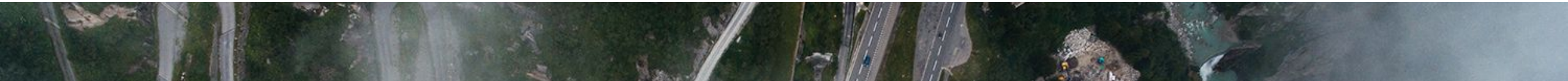
BRAINBOX

Report by Brainbox for the Global Partnership on AI

GIFCT Legal Frameworks Working Group

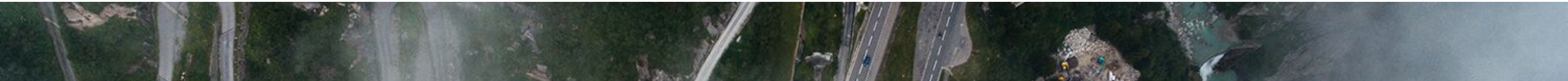
9 February 2022

www.brainbox.institute



About Brainbox

- LLC founded 2018 (NZ)
- Specialists in law and public policy consulting and research.
- Project funding: mixed public interest and commercial.
- Numerous projects across human rights, online expression, content moderation, emerging media technologies, computational law, access to justice, and health.



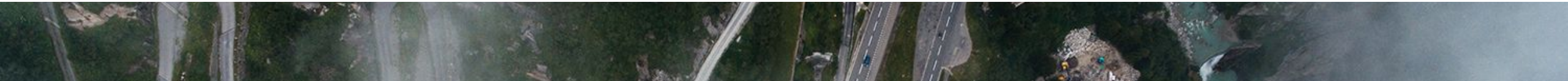
Report to the Social Media Investor Collaboration

BRAINBOX

- Led by New Zealand Super Fund, Neuberger Berman, Northern Trust
- Open letter to the Platforms
- 105 international investors
- Principles of responsible investment (won the UNPRI Stewardship award)
- Approx $13T under management
- GIFCT featured heavily in our analysis and recommendations

> “Are the changes made by Facebook, Twitter, and Google since the Christchurch livestream enough to prevent future harmful online content crises?”

www.brainbox.institute/investor-coalition-and-social-media

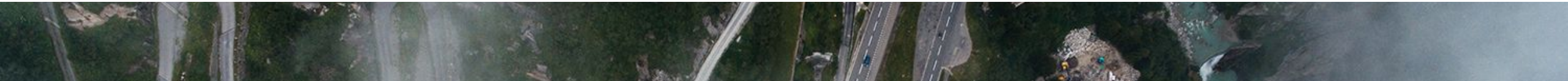


Follow-up on Content Moderation Regulation

- New Zealand Super Fund, Neuberger Berman, Northern Trust
- To strengthen their understanding of the regulatory area
- To help determine which regulation they should advocate for as responsible investors

What are the features, principles, and approaches for good content moderation regulation? Identify whether any existing regulation or proposals fit this.

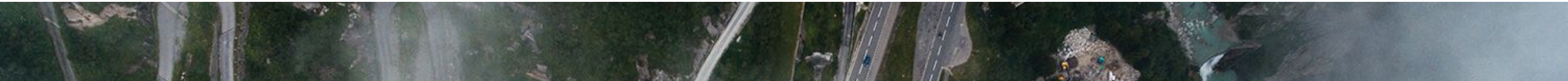
www.brainbox.institute/investor-coalition-and-social-media



Approach from GPAI

- Research proposal from Global Partnership on AI, funding from University of Otago
- Responsible AI working group
- “Fact-finding exercise” to embed researchers within Platforms and study “the effects of recommender systems”
- Brainbox was to identify and analyse key legal and policy issues to be anticipated in the GPAI research proposal

www.brainbox.institute/gpaiproposal

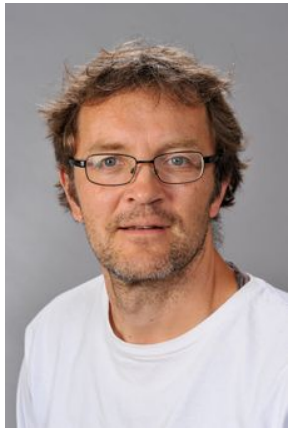


Global Partnership on AI

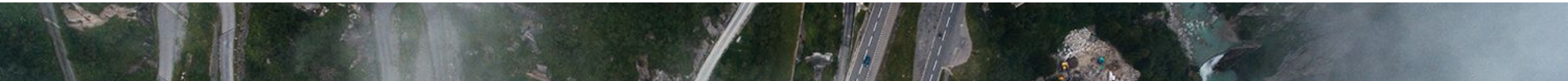
"...a **multistakeholder initiative** bringing together leading experts from science, industry, civil society, international organizations and government that share values to bridge the gap between theory and practice on AI by supporting cutting-edge research and applied activities on AI-related priorities.

We aim to provide a mechanism for **sharing multidisciplinary research and identifying key issues** among AI practitioners, with the objective of facilitating international collaboration, reducing duplication, acting as a global reference point for specific AI issues, and **ultimately promoting trust in and the adoption of trustworthy AI**. …"

(GPAI website)

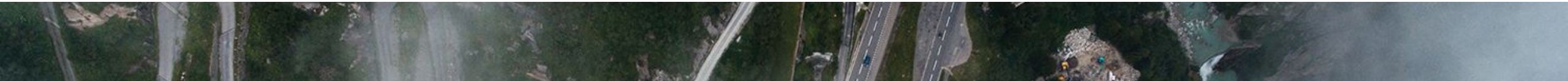


Professor Alistair Knott, University of Otago and GPAI



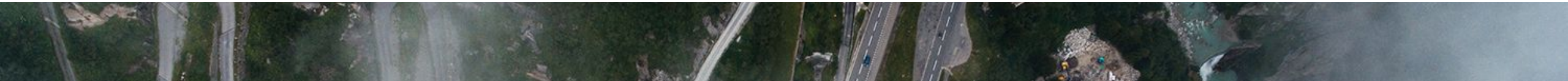
Specifics of GPAI proposal

- Proposal to engage in collaborative study with a social media platform
- Precise methods would be co-designed with the platform
- A “fact-finding exercise” to study whether recommender systems influence users over time toward TVEC. Includes literature review.
- Once agreed, methods would be audited by an embedded researcher
- Embedded researcher would be subject to legal obligations of confidentiality, etc
- More detail in technical report from GPAI group



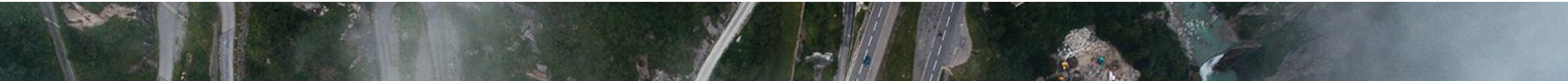
Context for our report

- Brainbox avoided picking a position on:
  - whether recommender systems cause harm
  - what content is harmful
- Law and public policy, not technical experts on AI methods
- Precise methods not laid out yet, method subject to change
- Collaboration is voluntary not mandatory, minimising legal issues caused by compulsion
- Method involves embedded researcher subject to private legal obligations of confidentiality, etc
- Jurisdictional variations across privacy, IP, confidentiality, etc



Purpose of Brainbox report

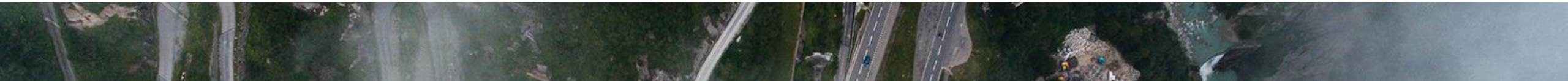
1. Identify and analyse law and policy issues that GPAI should account for in the design of its research
2. Anticipate the likely objections from platforms and other stakeholder groups (including GIFCT and human rights groups)
3. Assess grounds for these objections and whether they are justified
4. Recommend areas that deserve further investigation and thought



BRAINBOX

Selected issues

Report is 46 pages with extensive bibliography



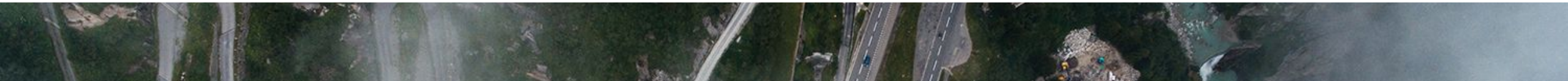
Executive summary

Transparency based approaches do create legal, regulatory, reputational, financial and commercial risks.

Risks of perception can be damaging whether or not they are justified.

"...there are compelling reasons why external study of recommender systems is so difficult. That is because access to information about recommender systems creates risk from a legal, regulatory, reputational, financial, and commercial perspective."

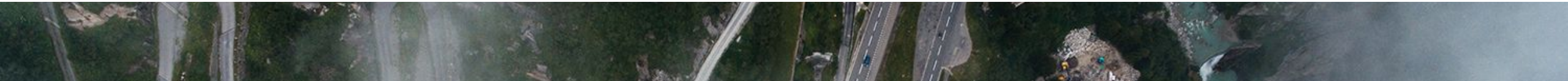
"It is also important to note that companies' resistance to external scrutiny of their recommender systems need not be explained solely by hostile or antisocial intent. There are a range of structural factors which may make it impossible for platforms to participate in this kind of research, even if they otherwise wished to participate. In particular, there is a complex network of rights and obligations that exists among regulators, users, platforms, contractors, and employees."



First class of issues

Adopting TVEC as a subject raises a host of difficult issues

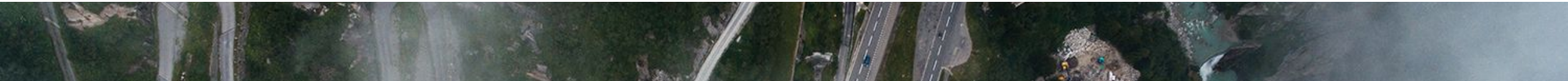
a. Applying a definition can be harder than articulating it
b. Natural language definitions
c. Operationalising natural language definitions in automated systems
d. Accuracy and reliability of automated systems
e. Legal consequences of labelling content as TVEC for platforms, users
f. Any “TVEC” on platforms already removed once identified
g. **Suggestion by GPAI to use hash sharing database as ML training set to create definition**
h. **“TVEC-adjacent” content - justification?**
i. **Predicting terrorist behaviour**



Second class of issues

Collaborative approach to external study

a. Usual issues around user privacy
b. Intellectual property and commercial confidentiality risks from disclosing details of business systems
c. Assurance and integrity issues if system parameters disclosed
d. Rights of third parties who are not users or employees (eg contractors, consultants)
e. Largely nullified by decision to embed researcher subject to legal obligations

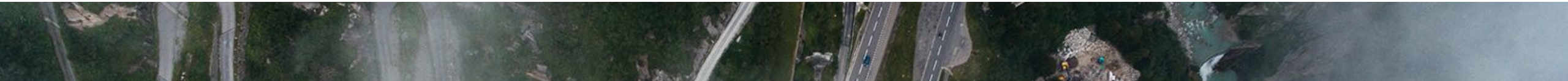


BRAINBOX

Two approaches to studying recommenders

1. Voluntary
   a. There are a number of these already
   b. Have apparently not satisfied expectations
   c. Should be studied closely by regulators and other transparency advocates
   d. Because they raise risks, always likely to be subject to controls over publication. That risks undermining perceived integrity.

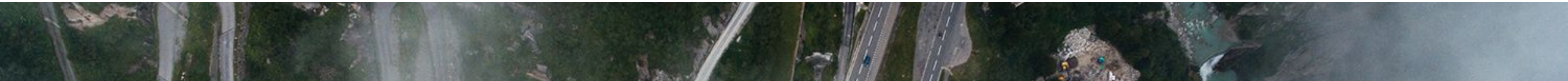
2. Mandated (by law)
   a. Existing network of rights and obligations must be re-organised (limited)
   b. We should acknowledge this directly and focus on substantive justification of these limitations



Voluntary approaches will give way to mandatory ones

BRAINBOX

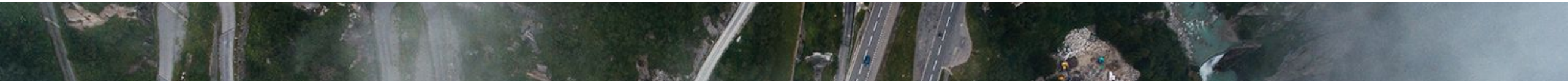
- Veto rights
- Perceptions of bad/good faith
- Difficulty complying with imprecise or over-broad access requests
- Delays
- Risks to companies and others
- Broad array of legal issues across multiple jurisdictions
- EU Digital Services Act “vetted researcher” regime



Reconfiguring rights and obligations requires human rights approach

BRAINBOX

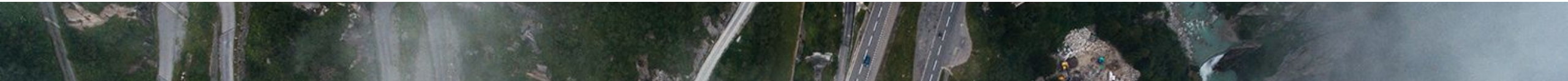
- Be up front about the fact that transparency approaches limit rights and interests of various actors (including platforms)
- Be led by principles of legality, justifiability, necessity, proportionality
- Demonstrable justification:
  - concerns about “TVEC-adjacent” content
  - “TVEC” already removed when identified
  - Identification is unavoidably difficult
- **Real concerns about delegating access arrangements to multi-stakeholder groups without sufficient legislative direction**



Platforms’ systems are integrated. Are narrow requests possible?

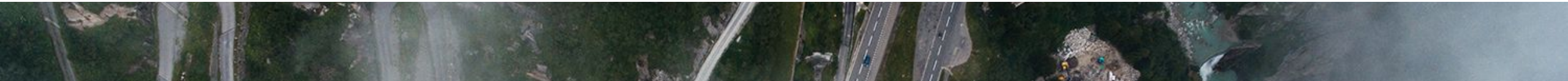
BRAINBOX

- Snowballing requests
- Hard to make requests without knowledge of systems
- Hard to respond to unclear requests
- Recommender systems moderate content
- Content moderation probably influences recommender systems
- Content moderation includes human operational processes
- Potential to link back to how user preferences are assessed and how all content is classified



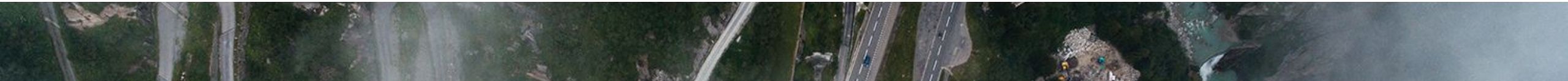
Conclusions

- Transparency based approaches, whether voluntary or mandatory:
  - do create risks to the companies.
  - can involve relinquishing existing legal rights.
- The benefit of the GPAI proposal is that it is being co-created with a view to minimising or avoiding these risks.
- Over time, voluntary arrangements will probably fail and become mandatory, because there will be irreconcilable disagreements.



Conclusions

- Mandatory arrangements must be specific and legislatures must provide clear direction about how to manage trade-offs
- A human rights approach should be preferred given the novelty of the legal and political issues involved
- Significant difficulty in framing access requests/regimes in ways that do not “snowball”
  - recommender ⟶ content moderation ⟶ operational human processes ⟶ core platform IP and commercial information
- Concern that Digital Services Act has not anticipated this difficulty or provided adequate direction.



BRAINBOX

Thank you

We welcome further discussion

www.brainbox.institute/projects

info@brainbox.institute