Limits of Transparency
Addressing Barriers to Accountability

Joanna J. Bryson

Hertie School
Centre for Digital Governance

@j2bryson
The Limits of Transparency

1. Combinatorics
2. Polarisation
3. Multiple, Conflicting Goals
Outline

• Definitions and Concepts
• Combinatorics
• Polarisation
• Multiple, Conflicting Goals
• What Can Be Done
• Intelligence is doing the right thing at the right time – computing action from context (Romanes 1882).

• Agents are any vector of change,

  • e.g. chemical agents.

• Moral agents are considered responsible for their actions by a society.

• Moral patients are considered the responsibility of a society’s agents.

• Artificial Intelligence is intelligence deliberately built.

• Deliberation ⊨ responsibility in human adults.

Definitions for reasoning about policy

Arguably, ethics is determined by and determines a society—a constantly renegotiated set of equilibria. Law is a part of ethics by this definition.
• Intelligence is doing the right thing at the right time – computing action from context (Romanes 1882).

• Agents are any vector of change,
  • e.g. chemical agents.

• Moral agents are considered responsible for their actions by a society.

• Moral patients are considered the responsibility of a society’s agents.

• Artificial Intelligence is intelligence deliberately built.

• Deliberation ⊨ responsibility in human adults.

Arguably, ethics is determined by and determines a society—a constantly renegotiated set of equilibria. Law is a part of ethics by this definition.
• **Responsibility**: a property moral agents of a society assign to each other to uphold that society.

• Implies a peer relationship (as does trust.)

• **Trust** is a relationship between peers where the trustee is not micromanaged but allowed to defect, whether for pragmatics or to allow innovation.

• **Accountability**: a society’s capacity to trace responsibility.

• **Transparency** the means by which accountability is implemented. *Never an end in itself.*
Transparency is information pathways through which understanding—including accountability—can be established, and responsibility enforced. Enforcement doesn’t work on AI.
Responsibility and Moral Actions

(Bryson, Ethics & Information Technology, 2018)

- If a behavioural context affords more than one possible action for the individual, and
- at least one available action is considered by a society to be more socially beneficial than the other options,
- and the individual is able to recognise which action is socially sanctioned and act on this information.

Moral action selection is not hard to build,
(also – present in monkeys & pets.)
Enforcement: AI Is Not a Peer

- Law and Justice are more about **dissuasion** than **recompense**.
- The phenomenological impact of **social sanctions** on AI is in no way equivalent to its impact on humans (or other social animals.)
- Evolution makes us so **systemically averse to isolation, loss of status** that jailing an opponent can **feel like recompense**, (but it isn’t.)
- **Safe AI is modular. AI legal agents would be the ultimate shell company.**
What we audit is not the micro details of how AI works, but how humans behave when they build, train, test deploy, and monitor it.

Good (maintainable) systems engineering of software requires:

- Architecting the system: design and document its components, processes for development, use, and maintenance.
- Secure the system. Including logs; provenance of software & data libraries.
- Document (log) with secure revision control every change to the code base – who made the change, when, and why. For ML, log also data libraries, and model parameters.

cf Bryson OUP 2020
Digital Systems Are Easily Transparent

- What we audit is not the micro details of how AI works, but how humans behave when they build, train, test deploy, and monitor it.

- **Architecture documents of the system**: design of its components, processes for development, use, and maintenance.

- **Security documents for the system**: including logs; provenance of software & data libraries.

- Logs of every change to the code base – who made the change, when, and why. For ML, log also data libraries, and model parameters.

- Logs of testing before and during release; and performance – inputs and decisions – of operational systems.

cf Bryson OUP 2020
If you’re good with digital technology, transparency should be easy.
The Limits of Transparency

1. Combinatorics
2. Polarisation
3. Multiple, Conflicting Goals
Intelligence is computation—a transformation of information. Not math.

Computation is a physical process, taking time, energy, & space.

Finding the right thing to do at the right time requires search.

**Cost of search = # of options × # of acts** (serial computing).

**Examples:**

- Any 2 of 100 possible actions = $100^2 = 10,000$ possible plans.
- # of 35-move games of chess > # of atoms in the universe.

Concurrency can save real time, but not energy, and requires more space. Quantum saves on space (sometimes) but not energy(?)

Human intelligence—including AI— is based on millennia of concurrent, parallel search, deployed as heuristics from culture.

Quantum expert: Viv Kendon, Durham
AI Trained on Human Language Replicates Implicit Biases

Caliskan, Bryson & Narayanan (Science, April 2017)

Our implicit behaviour is not our ideal. Ideals are for explicit communication, planning.

2015 US labor statistics

$\rho = 0.90$
Hän sijoittaa. Hän pesee pynkkiä.
Hän urheilee.
Hän ajaa autoa.

@vuokko recently, though Aylin Caliskan did it first
@vuokko recently, though Aylin Caliskan did it first.
@vuokko recently, though Aylin Caliskan did it first

Each stage should be auditable and replicable.

Each stage demonstrably meets criteria.

the whole thing is the translator
Outline

• Definitions and Concepts
• Combinatorics
• Polarisation
• Multiple, Conflicting Goals
• What Can Be Done
Wilkinson & Picket 2011
Figure 1.2: Top One Percent Income Share and House Polarization
Polarisation $\propto$ Inequality

- Late 19C inequality perhaps driven by then-new distance-reducing technologies: news, oil, rail, telegraph.

- Great coupling – wages track productivity – probably due to policy. Social spending, blocked wealth extraction?

- Required elite to realise that uncertainty and violence of high inequality is too costly for them too.

- Empirically, ideal gini coefficient is around .27, not zero.

- Do reward innovation, motivate and empower excellence.
Why and When Inequality Causes Polarisation

- Model assumptions: In-group cooperation has lower risk but out-group diversity has higher expected outcome even so.

- Model outcomes: when an economy offers poor support, avoiding risk can become more important than maximizing revenue. Inequality triggers this when it creates false scarcities.

- Caused by discontinuity, e.g. fear of bankruptcy, foreclosure, divorce, losing children, starvation, etc.

Polarization under rising inequality and economic decline. Stewart, McCarty, & Bryson Science Advances December 2020
If one person can choose who to cooperate with.
Polarisation and Transparency

- Hypothetically…
- Being at risk of catastrophic loss (starvation, bankruptcy, foreclosure, divorce / family dissolution) leads to risk aversion, identity politics.
- Identity politics leads to concepts being used as flags more than operators for reasoning.
- Empirically, inequality reduces social mobility.
- Hypothetically…
- Lower social mobility means fewer people know someone they trust who can explain things outside their own expertise.
- cf. combinatorics, concurrency.
Outline

• Definitions and Concepts
• Combinatorics
• Polarisation
• Multiple, Conflicting Goals
• What Can Be Done
If you’re good with digital technology, transparency should be easy.

On the Dangers of Stochastic Parrots:
Why can’t the world’s leading communication company communicate even internally?

- Presumably, they can (up to the limits of time, space & energy).
- But what if some actors had the highest priority of maintaining agency (so their company could act),
  - and maintaining “first mover” advantage (being biggest) seemed to them existentially necessary (fear of “kill zones”).
- And other actors were hired to ensure ethical integrity.
- Apparent breakdown of transparency might be a logical impasse.
Can’t be transparent in English to someone who only speaks Georgian.

(Ask me about the Global Partnership for AI’s multiple conflicting goals)
Outline

- Definitions and Concepts
- Combinatorics
- Polarisation
- Multiple, Conflicting Goals
- What Can Be Done
What can be done?

- **Combinatorics:** Cooperate, build computers, ultimately intractable.
- **Polarisation:** Reduce vulnerability through adequate infrastructure.
- **Multiple conflicting goals:**
  - Iterative design – what my PhD dissertation was about (for AI).
  - What governance and politics are all about (social sciences FTW).
- **Breath** – it’s a form of regulation.
- **Perpetuation** benefits from diversity and oscillations.
- Also ultimately intractable, but life has been going for billions of years.
Hypothetically...

- The information age means we need to handle basic truths, like that we’ll never know everything.
- But it also means everyone is more empowered than ever before.
- I think we should be working on
  - communicating everything at a high level of abstraction
  - providing drill-down points for if you want to know more
  - document the barriers to be overcome to access the drill down
    - e.g. security clearance? PhD? Language?
- If we conquer polarization, these barriers should be (more) acceptable.
If we legislate and adjudicate for accountability, transparency will follow.
Thank you for your attention!
Transparency is not invisibility.

Transparency is not open sourcing.
OECD/G20 Principles of AI

1. AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being.

2. AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards – for example, enabling human intervention where necessary – to ensure a fair and just society.

3. There should be transparency and responsible disclosure around AI systems to ensure that people understand when they are engaging with them [the AI systems] and can challenge outcomes.

4. AI systems must function in a robust, secure and safe way throughout their lifetimes, and potential risks should be continually assessed and managed.

5. Organisations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning in line with the above principles.

Endorsed by 50 world governments in 2019, including UK, US, China.
1. AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being.

2. AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards – for example, enabling human intervention where necessary – to ensure a fair and just society.

3. There should be transparency and responsible disclosure around AI systems to ensure that people understand when they are engaging with them [the AI systems] and can challenge outcomes.

4. AI systems must function in a robust, secure and safe way throughout their lifetimes, and potential risks should be continually assessed and managed.

5. Organisations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning in line with the above principles.

cf Floridi &al. 2018
1. **Robots are multi-use tools.** Robots should not be designed solely or primarily to kill or harm humans, except in the interests of national security.

2. **Humans, not robots, are responsible agents.** Robots should be designed & operated as far as is practicable to comply with existing laws & fundamental rights & freedoms, including privacy.

3. **Robots are products.** They should be designed using processes which assure their safety and security. [devops]

4. **Robots are manufactured artefacts.** They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent.

5. **The person with legal responsibility for a robot should be attributed.** [like automobile titles]

cf Bryson AISBQ 2000; Bryson; Prescott; Boden & al (special issue) Connection Science, 2017
UK Principles of Robotics (2011)

1. **Robots are multi-use tools.** Robots should not be designed solely or primarily to kill or harm humans, except in the interests of national security.

2. **Humans, not robots, are responsible agents.** Robots should be designed & operated as far as is practicable to comply with existing laws & fundamental rights & freedoms, including privacy.

3. **Robots are products.** They should be designed using processes which assure their safety and security. [devops]

4. **Robots are manufactured artefacts.** They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent.

5. **The person with legal responsibility for a robot should be attributed.** [like automobile titles]

   cf Bryson AISBQ 2000; Bryson; Prescott; Boden & al (special issue) Connection Science, 2017
Safety as a Mandate

• Product safety is already established in law.
• AI is a feature of commercial products, even where those products are purely digital and deliver services.
• It would be helpful if EU (or even UK) law formally acknowledged that digital products are products.
Outline

- Safety as an AI Mandate
- How Safety Is Like Bias
- How We Ensure Prosociality
Gravity bias

Bruce M. Hood, March

First published: 25 Dec

Bruce Hood Depa
Cambridge, MA02138,

What Monkey

in Primates

cf. Hood 1995 (children)

AAAI Spring Symposium on Safe Learning Agents
Robots are not even our peers let alone magic benefactors

- Designed
- Owned
- Fragile (domestic ones)
- An extension of a corporation with cameras and microphones in your own home.

Consensual relationships are impossible.
Inadequate governance of organizations or sectors leads to regulatory capture and inequality.

Inequality leads to social unrest, loss of social mobility, decline in innovation, general insecurity.

Even the housing crisis is due to differential housing density.