1. What can go wrong?

1. What can go wrong?
2. How do algorithms make things worse?

1. What can go wrong?
2. How do algorithms make things worse?
3. Can regulation help?

What can go wrong? (1 of 3)

What can go wrong? (1 of 3)

What can go wrong? (1 of 3)

VIOLENCE

BAD DECISIONS

What can go wrong? (1 of 3)

VIOLENCE

BAD DECISIONS

FREEDOM OF EXPRESSION

POLARISATION

RADICALISATION

ONLINE HARMS

What can go wrong? (1 of 3)

VIOLENCE

BAD DECISIONS

FREEDOM OF EXPRESSION

POLARISATION

RADICALISATION

ONLINE HARMS

EPISTEMIC INJUSTICE

POLARISATION

AUTONOMY

PRIVACY

TRUST

What can go wrong? (1 of 3)

VIOLENCE

BAD DECISIONS

FREEDOM OF EXPRESSION

POLARISATION

RADICALISATION

ONLINE HARMS

EPISTEMIC INJUSTICE

POLARISATION

AUTONOMY

PRIVACY

TRUST

INFORMATION HAZARDS

ATTENTION SPANS

INFLUENCE OPERATIONS

NEWS DESERTS

How do algorithms make things worse? (2 of 3)

How do algorithms make things worse? (2 of 3)

NOT “ALIGNED”

“ALIGNED”

How do algorithms make things worse? (2 of 3)

NOT “ALIGNED”

PERFORMS TASK IMPERFECTLY

“ALIGNED”

How do algorithms make things worse? (2 of 3)

NOT “ALIGNED”

PERFORMS TASK IMPERFECTLY

- Content Moderation

“ALIGNED”

How do algorithms make things worse? (2 of 3)

NOT “ALIGNED”

PERFORMS TASK IMPERFECTLY

- Content Moderation

“ALIGNED”

> [A] probabilistic conception of online speech acknowledges that enforcement of the rules made as a result of this balancing will never be perfect, and so governance systems **should take into account the inevitability of error** and **choose what kinds of errors to prefer**. The conscious acceptance of the fact that getting speech determinations wrong in some percentage of cases is inherent in online speech governance requires being **much more candid about error rates**, which can allow for the **calibration of rulemaking to the practical realities of enforcement**.
>
> — evelyn douek, *Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability* (2021)

How do algorithms make things worse? (2 of 3)

NOT “ALIGNED”

PERFORMS TASK IMPERFECTLY

- Content Moderation

WE SET THE WRONG TASK / TASK IS SUBJECTIVE

“ALIGNED”

How do algorithms make things worse? (2 of 3)

NOT “ALIGNED”

PERFORMS TASK IMPERFECTLY

- Content Moderation

WE SET THE WRONG TASK / TASK IS SUBJECTIVE

- Content Moderation, again (who to surveil? what to suppress?, debate further fuels animosity)

“ALIGNED”

How do algorithms make things worse? (2 of 3)

NOT “ALIGNED”

“ALIGNED”

Proportionality **necessitates intrusions on rights being justified**, and greater intrusions have stronger justifications. In constitutional systems, proportionality takes various doctrinal forms but always involves a balancing test that requires the decisionmaker to balance societal interests against individual rights. This emphasis on justification and balancing therefore **takes the decisionmaker from being a mere "taxonomist" (categorizing types of content) to grocer (placing competing interests on a scale and weighing them against each other)**. This task requires much greater transparency of reasoning.

— evelyn douek, *Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability* (2021)

How do algorithms make things worse? (2 of 3)

NOT “ALIGNED”

“ALIGNED”

Proportionality **necessitates intrusions on rights being justified**, and greater intrusions have stronger justifications. In constitutional systems, proportionality takes various doctrinal forms but always involves a balancing test that requires the decisionmaker to balance societal interests against individual rights. This emphasis on justification and balancing therefore **takes the decisionmaker from being a mere "taxonomist" (categorizing types of content) to grocer (placing competing interests on a scale and weighing them against each other)**. This task requires much greater transparency of reasoning.

— evelyn douek, *Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability* (2021)

How do algorithms make things worse? (2 of 3)

NOT “ALIGNED”

PERFORMS TASK IMPERFECTLY

- Content Moderation

WE SET THE WRONG TASK / TASK IS SUBJECTIVE

- Content Moderation, again (who to surveil? what to suppress?, debate further fuels animosity)

- Engagement Optimisation (basis in *revealed preference*, metaphors, information cascades, attention hijacking)

“ALIGNED”

How do algorithms make things worse? (2 of 3)

NOT “ALIGNED”

PERFORMS TASK IMPERFECTLY

- Content Moderation

WE SET THE WRONG TASK / TASK IS SUBJECTIVE

- Content Moderation, again (who to surveil? what to suppress?, debate further fuels animosity)
- Engagement Optimisation (basis in *revealed preference*, metaphors, information cascades, attention hijacking)

PERFORMS TASK, BUT HAS SIDE EFFECTS

“ALIGNED”

How do algorithms make things worse? (2 of 3)

NOT "ALIGNED"

PERFORMS TASK IMPERFECTLY

- Content Moderation

WE SET THE WRONG TASK / TASK IS SUBJECTIVE

- Content Moderation, again (who to surveil? what to suppress?, debate further fuels animosity)

- Engagement Optimisation (basis in *revealed preference*, metaphors, information cascades, attention hijacking)

PERFORMS TASK, BUT HAS SIDE EFFECTS

- User Tampering (e.g. RL or AB-tested recommenders)

"ALIGNED"

How do algorithms make things worse? (2 of 3)

NOT "ALIGNED"

PERFORMS TASK IMPERFECTLY

- Content Moderation

WE SET THE WRONG TASK / TASK IS SUBJECTIVE

- Content Moderation, again (…eil? what to suppress?, …er fuels animosity)
- Engagement Optimisation (bas… *revealed preference*, …s, information cascades, attention hijacking)

PERFORMS TASK, BUT HAS SIDE EFFECTS

- User Tampering (e.g. RL or AB-tested recommenders)

AGENT

"ALIGNED"

ENVIRONMENT

How do algorithms make things worse? (2 of 3)

NOT "ALIGNED"

PERFORMS TASK IMPERFECTLY

- Content Moderation

WE SET THE WRONG TASK / TASK IS SUBJECTIVE

- Content Moderation, again (…eil? what to suppress?, …er fuels animosity)
- Engagement Optimisation (bas… *revealed preference*, …s, information cascades, attention hijacking)

PERFORMS TASK, BUT HAS SIDE EFFECTS

- User Tampering (e.g. RL or AB-tested recommenders)

AGENT

"ALIGNED"

ENVIRONMENT

How do algorithms make things worse? (2 of 3)

NOT "ALIGNED"

PERFORMS TASK IMPERFECTLY

- Content Moderation

WE SET THE WRONG TASK / TASK IS SUBJECTIVE

- Content Moderation, again (who to surveil? what to suppress?, debate further fuels animosity)
- Engagement Optimisation (basis in *revealed preference*, metaphors, information cascades, attention hijacking)

PERFORMS TASK, BUT HAS SIDE EFFECTS

- User Tampering (e.g. RL or AB-tested recommenders)
- Cognitive Atrophy (e.g. automated fact-checking, AI research assistants)

"ALIGNED"

How do algorithms make things worse? (2 of 3)

NOT "ALIGNED"

PERFORMS TASK IMPERFECTLY

- Content Moderation

WE SET THE WRONG TAS

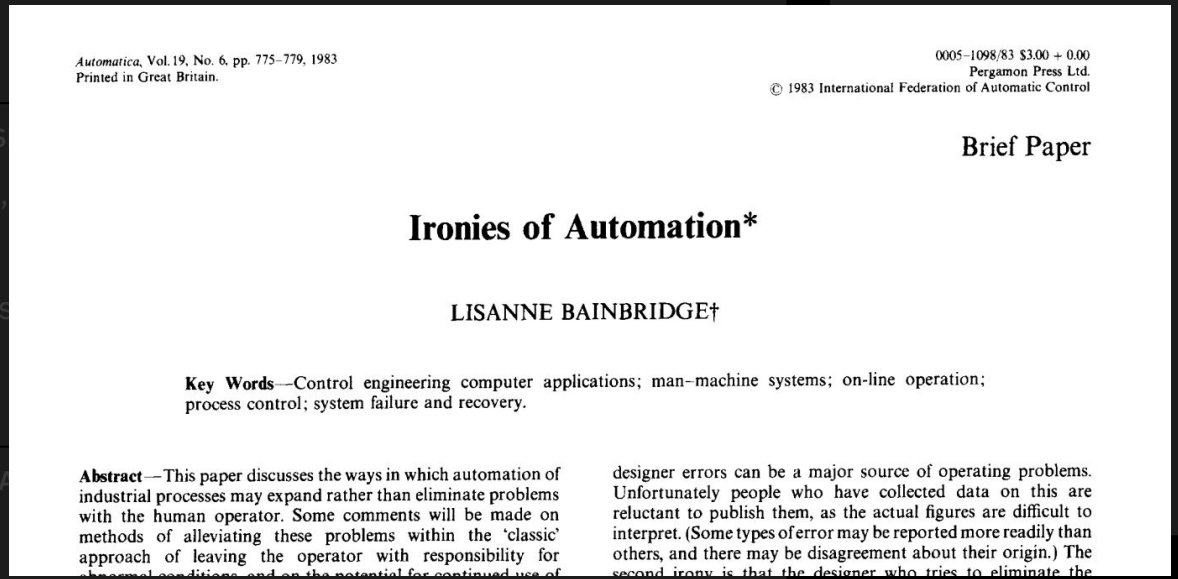
- Content Moderation,
- Engagement Optimis

PERFORMS TASK, BUT HA

- User Tampering
- Cognitive Atrophy

"ALIGNED"

(e.g. automated fact-checking, AI research assistants)



Automatica, Vol. 19, No. 6, pp. 775–779, 1983
Printed in Great Britain.

0005–1098/83 $3.00 + 0.00
Pergamon Press Ltd.

Brief Paper

Ironies of Automation*

LISANNE BAINBRIDGE†

Key Words—Control engineering computer applications; man–machine systems; on-line operation; process control; system failure and recovery.

Abstract—This paper discusses the ways in which automation of industrial processes may expand rather than eliminate problems with the human operator. Some comments will be made on methods of alleviating these problems within the 'classic' approach of leaving the operator with responsibility for

designer errors can be a major source of operating problems. Unfortunately people who have collected data on this are reluctant to publish them, as the actual figures are difficult to interpret. (Some types of error may be reported more readily than others, and there may be disagreement about their origin.) The

How do algorithms make things worse? (2 of 3)

NOT “ALIGNED”

PERFORMS TASK IMPERFECTLY

- Content Moderation

WE SET THE WRONG TASK / TASK IS SUBJECTIVE

- Content Moderation, again (who to surveil? what to suppress?, debate further fuels animosity)
- Engagement Optimisation (basis in *revealed preference*, metaphors, information cascades, attention hijacking)

PERFORMS TASK, BUT HAS SIDE EFFECTS

- User Tampering (e.g. RL or AB-tested recommenders)
- Cognitive Atrophy (e.g. automated fact-checking, AI research assistants)
- Perverse Incentives (e.g. financially-motivated misinfo)

“ALIGNED”

How do algorithms make things worse? (2 of 3)

NOT "ALIGNED"

PERFORMS TASK IMPERFECTLY

- Content Moderation

WE SET THE WRONG TASK / TASK IS SUBJECTIVE

- Content Moderation, again (who to surveil? what to suppress?, debate further fuels animosity)
- Engagement Optimisation (basis in *revealed preference*, metaphors, information cascades, attention hijacking)

PERFORMS TASK, BUT HAS SIDE EFFECTS

- User Tampering (e.g. RL or AB-tested recommenders)
- Cognitive Atrophy (e.g. automated fact-checking, AI research assistants)
- Perverse Incentives (e.g. financially-motivated misinfo)

"ALIGNED"

USED BY BAD ACTORS

How do algorithms make things worse? (2 of 3)

NOT "ALIGNED"

PERFORMS TASK IMPERFECTLY

- Content Moderation

WE SET THE WRONG TASK / TASK IS SUBJECTIVE

- Content Moderation, again (who to surveil? what to suppress?, debate further fuels animosity)
- Engagement Optimisation (basis in *revealed preference*, metaphors, information cascades, attention hijacking)

PERFORMS TASK, BUT HAS SIDE EFFECTS

- User Tampering (e.g. RL or AB-tested recommenders)
- Cognitive Atrophy (e.g. automated fact-checking, AI research assistants)
- Perverse Incentives (e.g. financially-motivated misinfo)

"ALIGNED"

USED BY BAD ACTORS

- Synthetic Media
  (deep fakes, spear phishing, etc.)

How do algorithms make things worse? (2 of 3)

NOT "ALIGNED"

PERFORMS TASK IMPERFECTLY

- Content Moderation

WE SET THE WRONG TASK / TASK IS SUBJECTIVE

- Content Moderation, again (who to surveil? what to suppress?, debate further fuels animosity)
- Engagement Optimisation (basis in *revealed preference*, metaphors, information cascades, attention hijacking)

PERFORMS TASK, BUT HAS SIDE EFFECTS

- User Tampering (e.g. RL or AB-tested recommenders)
- Cognitive Atrophy (e.g. automated fact-checking, AI research assistants)
- Perverse Incentives (e.g. financially-motivated misinfo)

"ALIGNED"

USED BY BAD ACTORS

- Synthetic Media (deep fakes, spear phishing, etc.)
- Persuasion Tools (ad tech, information gerrymandering)

How do algorithms make things worse? (2 of 3)

NOT "ALIGNED"

PERFORMS TASK IMPERFECTLY

- Content Moderation

WE SET THE WRONG TASK / TASK IS SUBJECTIVE

- Content Moderation, again (who to surveil? what to suppress?, debate further fuels animosity)
- Engagement Optimisation (basis in *revealed preference*, metaphors, information cascades, attention hijacking)

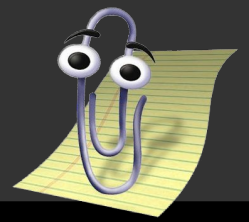
PERFORMS TASK, BUT HAS SIDE EFFECTS

- User Tampering (e.g. RL or AB-tested recommenders)
- Cognitive Atrophy (e.g. automated fact-checking, AI research assistants)
- Perverse Incentives (e.g. financially-motivated misinfo)

"ALIGNED"

USED BY BAD ACTORS

- Synthetic Media
  (deep fakes, spear phishing, etc.)
- Persuasion Tools
  (ad tech, information gerrymandering)



How do algorithms make things worse? (2 of 3)

NOT "ALIGNED"

PERFORMS TASK IMPERFECTLY

- Content Moderation

WE SET THE WRONG TASK / TASK IS SUBJECTIVE

- Content Moderation, again (who to surveil? what to suppress?, debate further fuels animosity)
- Engagement Optimisation (basis in *revealed preference*, metaphors, information cascades, attention hijacking)

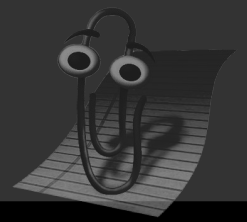
PERFORMS TASK, BUT HAS SIDE EFFECTS

- User Tampering (e.g. RL or AB-tested recommenders)
- Cognitive Atrophy (e.g. automated fact-checking, AI research assistants)
- Perverse Incentives (e.g. financially-motivated misinfo)

"ALIGNED"

USED BY BAD ACTORS

- Synthetic Media
  (deep fakes, spear phishing, etc.)
- Persuasion Tools
  (ad tech, information gerrymandering)



NOTABLE EXCEPTIONS

How do algorithms make things worse? (2 of 3)

NOT "ALIGNED"

PERFORMS TASK IMPERFECTLY

- Content Moderation

WE SET THE WRONG TASK / TASK IS SUBJECTIVE

- Content Moderation, again (who to surveil? what to suppress?, debate further fuels animosity)
- Engagement Optimisation (basis in *revealed preference*, metaphors, information cascades, attention hijacking)

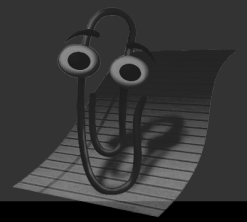
PERFORMS TASK, BUT HAS SIDE EFFECTS

- User Tampering (e.g. RL or AB-tested recommenders)
- Cognitive Atrophy (e.g. automated fact-checking, AI research assistants)
- Perverse Incentives (e.g. financially-motivated misinfo)

"ALIGNED"

USED BY BAD ACTORS

- Synthetic Media
  (deep fakes, spear phishing, etc.)
- Persuasion Tools
  (ad tech, information gerrymandering)



NOTABLE EXCEPTIONS

"filter bubbles" / "echo chambers"
(lack of evidence)

How do algorithms make things worse? (2 of 3)

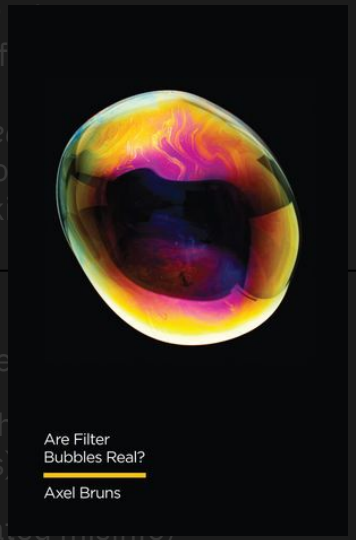
NOT "ALIGNED"

PERFORMS TASK IMPERFECTLY

- Content Moderation

WE SET THE WRONG TASK / TASK IS SUBJECTIVE

- Content Moderation, again (who to surveil? debate further f
- Engagement Optimisation (basis in *reveale* metaphors, info attention hijack



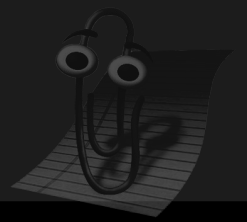
PERFORMS TASK, BUT HAS SIDE EFFECTS

- User Tampering (e.g. RL or AB-tested re
- Cognitive Atrophy (e.g. automated fact-ch AI research assistants)
- Perverse Incentives (e.g. financially-motivated [illegible])

"ALIGNED"

USED BY BAD ACTORS

- Synthetic Media (deep fakes, spear phishing, etc.)
- Persuasion Tools (ad tech, information gerrymandering)



NOTABLE EXCEPTIONS

"filter bubbles" / "echo chambers" (lack of evidence)

Can regulation help? (3 of 3)

Can regulation help? (3 of 3)

CHALLENGES

Can regulation help? (3 of 3)

CHALLENGES

- Reactiveness

Can regulation help? (3 of 3)

CHALLENGES

- Reactiveness
- Lack of good options

Can regulation help? (3 of 3)

CHALLENGES

- Reactiveness
- Lack of good options
- Free expression

Can regulation help? (3 of 3)

CHALLENGES

- Reactiveness
- Lack of good options
- Free expression
- Enforcement? Brandolini’s law

Can regulation help? (3 of 3)

CHALLENGES

- Reactiveness
- Lack of good options
- Free expression
- Enforcement? Brandolini’s law

POSSIBLE REGULATORY MODELS

Can regulation help? (3 of 3)

CHALLENGES

- Reactiveness
- Lack of good options
- Free expression
- Enforcement? Brandolini's law

POSSIBLE REGULATORY MODELS

- Transparency in platforms
  - Data Access Centres
  - A/B effect

Can regulation help? (3 of 3)

CHALLENGES

- Reactiveness
- Lack of good options
- Free expression
- Enforcement? Brandolini's law

POSSIBLE REGULATORY MODELS

- Transparency in platforms
  - Data Access Centres
  - A/B effect
- Friction

Can regulation help? (3 of 3)

CHALLENGES

- Reactiveness
- Lack of good options
- Free expression
- Enforcement? Brandolini's law

POSSIBLE REGULATORY MODELS

- Transparency in platforms
  - Data Access Centres
  - A/B effect
- Friction
- Interoperability, middleware (cf. Fukuyama)

Can regulation help? (3 of 3)

CHALLENGES

- Reactiveness
- Lack of good options
- Free expression
- Enforcement? Brandolini's law

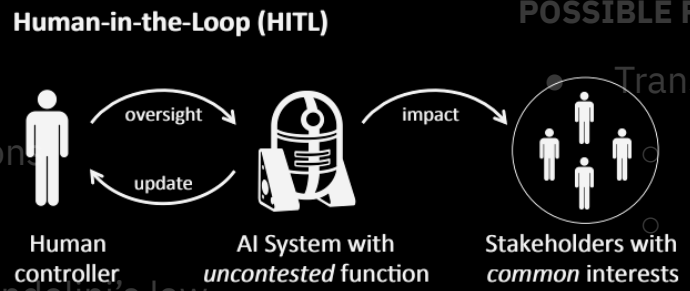
POSSIBLE REGULATORY MODELS

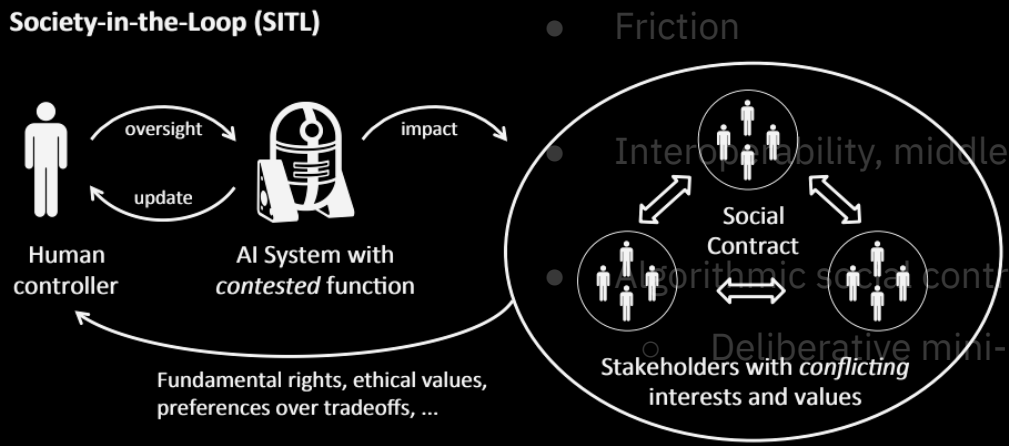
- Transparency in platforms
  - Data Access Centres
  - A/B effect
- Friction
- Interoperability, middleware (cf. Fukuyama)
- Algorithmic social contract
  - Deliberative mini-publics (cf. Ovadya)

Can regulation help? (3 of 3)

CHALLENGES

- Reactiveness
- Lack of good options
- Free expression
- Enforcement? Brandolini's law





POSSIBLE REGULATORY MODELS

- Transparency in platforms
  - Data Access Centres
  - A/B effect
- Friction
- Interoperability, middleware (cf. Fukuyama)
- Algorithmic social contract
  - Deliberative mini-publics (cf. Ovadya)

— Iyad Rahwan, *Society-in-the-Loop: Programming the Algorithmic Social Contract* (2017)

Can regulation help? (3 of 3)

CHALLENGES

- Reactiveness
- Lack of good options
- Free expression
- Enforcement? Brandolini's law

POSSIBLE REGULATORY MODELS

- Transparency in platforms
  - Data Access Centres
  - A/B effect
- Friction
- Interoperability, middleware (cf. Fukuyama)
- Algorithmic social contract
  - Deliberative mini-publics (cf. Ovadya)

Can regulation help? (3 of 3)

CHALLENGES

- Reactiveness
- Lack of good options
- Free expression
- Enforcement? Brandolini's law

POSSIBLE REGULATORY MODELS

- Transparency in platforms
  - Data Access Centres
  - A/B effect
- Friction
- Interoperability, middleware (cf. Fukuyama)
- Algorithmic social contract
  - Deliberative mini-publics (cf. Ovadya)

good regulation of AI has meta-benefits