How to Assess Trustworthy AI *with* Z-inspection®

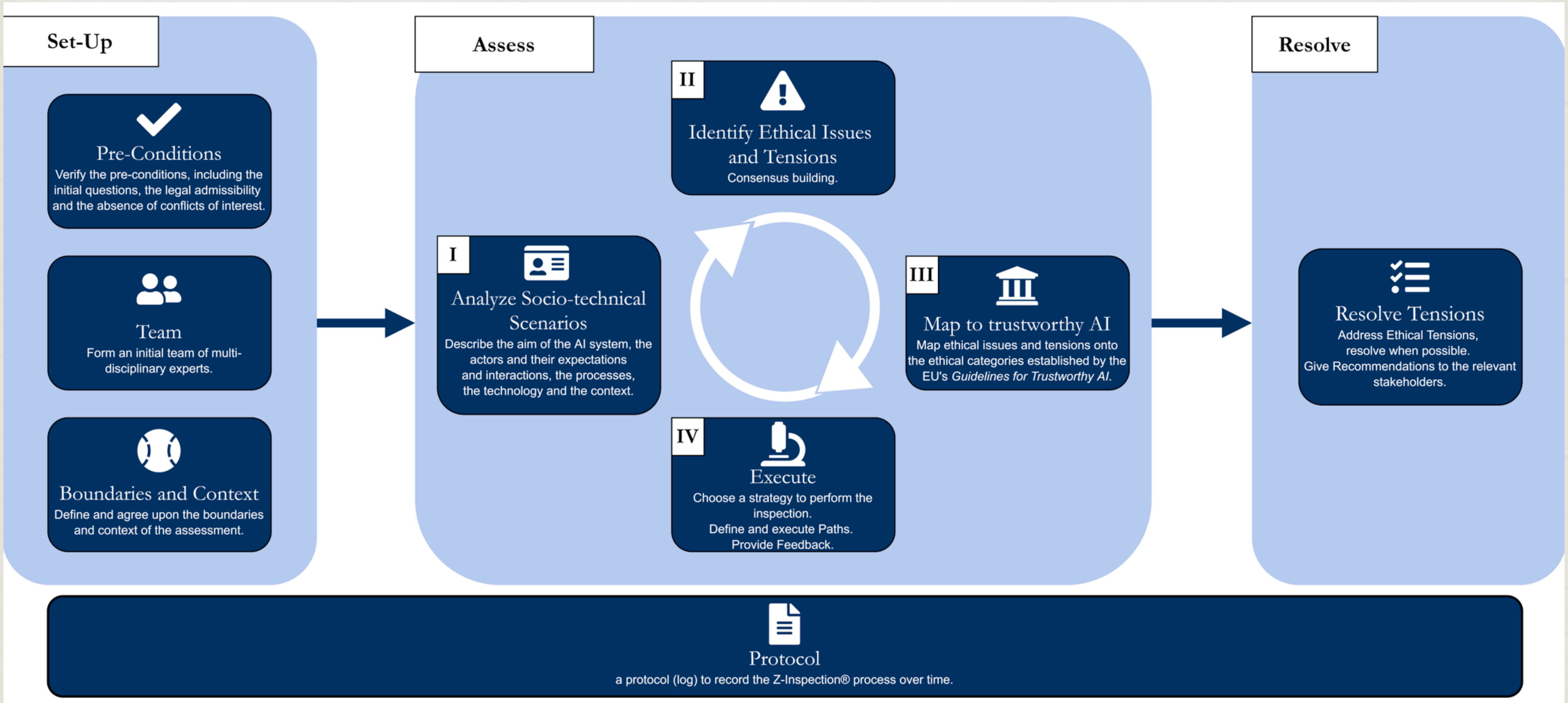
Roberto V. Zicari

Z-Inspection® Initiative

http://z-inspection.org

Graduate School of Data Science, SNU, Seoul

Z-inspection® Process in a Nutshell



Set-Up

Pre-Conditions
Verify the pre-conditions, including the initial questions, the legal admissibility and the absence of conflicts of interest.

Team
Form an initial team of multi-disciplinary experts.

Boundaries and Context
Define and agree upon the boundaries and context of the assessment.

Assess

I
Analyze Socio-technical Scenarios
Describe the aim of the AI system, the actors and their expectations and interactions, the processes, the technology and the context.

II
Identify Ethical Issues and Tensions
Consensus building.

III
Map to trustworthy AI
Map ethical issues and tensions onto the ethical categories established by the EU's *Guidelines for Trustworthy AI.*

IV
Execute
Choose a strategy to perform the inspection.
Define and execute Paths.
Provide Feedback.

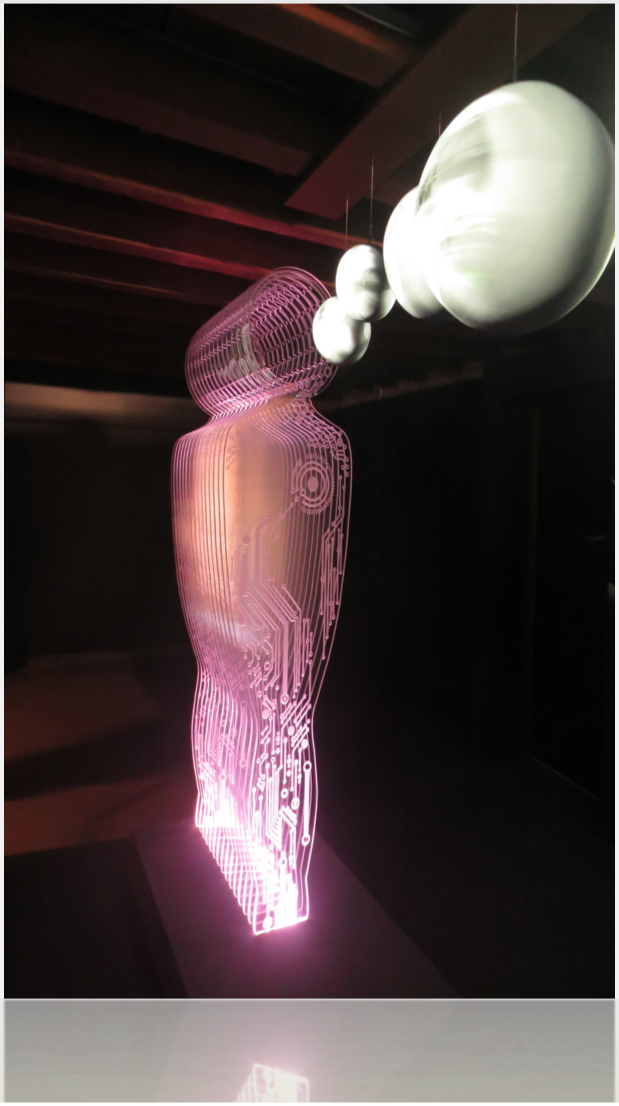
Resolve

Resolve Tensions
Address Ethical Tensions, resolve when possible.
Give Recommendations to the relevant stakeholders.

Protocol
a protocol (log) to record the Z-Inspection® process over time.

Ethical Tensions

(photo RVZ,Venice Biennale)



Ethical Tensions

- "We use the umbrella term 'tension' to refer to different ways in which values can be in conflict, some more fundamentally than others."
- "When we talk about tensions between values, we mean tensions between the pursuit **of different values in technological applications** rather than an abstract tension between the values themselves."

Source: *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Identify Tensions and Trade-offs

- **Tensions** may arise between ethical principles**, for which there is no fixed solution**.

- “In line with the EU fundamental commitment to democratic engagement, due process and open political participation, *methods of accountable deliberation to deal with such tensions should be established.* “

source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019

We use a Catalog of predefined ethical tensions

- To help the process, especially as a help to experts who might have not sufficient knowledge in ethics, we used a sample of catalog of predefined ethical tensions.

- We have chosen the catalog defined by the Nuffield Foundations (Whittlestone et al., 2019)

- Source: *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Catalogue of Tensions

- *Accuracy vs. Fairness*
- *Accuracy vs. Explainability*
- *Privacy vs. Transparency*
- *Quality of services vs. Privacy*
- *Personalisation vs. Solidarity*
- *Convenience vs. Dignity*
- *Efficiency vs. Safety and Sustainability*
- *Satisfaction of Preferences vs. Equality*

[1] Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Catalogue of Tensions

- **Accuracy** *versus* **Fairness**:

“An AI system which is most accurate on average may systematically discriminate against a specific minority.”

Tension

Using AI systems to make decisions and predictions more ***accurate*** versus ***ensuring fair and equal treatment***.

Hypothetical illustration

- "To assist in decisions about whether to release defendants *on bail* (*) or *to grant parole* (**), a jurisdiction adopts an AI system that estimates the 'recidivism risk' of criminal defendants, i.e. their likelihood of re-offending.

- (*) **Bail** is the conditional release of a defendant with the promise to appear in court when required.

- (**) If a prisoner is given **parole**, he or she is released before the official end of their prison sentence and has to promise to behave well.

Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Hypothetical illustration

- “Although it is **highly accurate on average, it systematically discriminates against black defendants**, because the ‘false positives’ – the rate of individuals classed as high risk who did not go on to reoffend – is almost twice as high for black as for white defendants.“

- Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Hypothetical illustration

- If “the inner workings of the AI system is a **trade secret** of the company that produced it (and in any case is too complex for any individual to understand), the defendants have little to no recourse to challenging the verdict that have huge consequences on their lives. “

Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Catalogue of Tensions

- **Accuracy *versus* Explainability**:

"the most accurate AI systems may be based on complex methods (such as deep learning), the internal logic of which its developers or users do not fully understand. "

- Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research*. Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London*. Nuffield Foundation.

Catalogue of Tensions

- **Quality of Services** *versus* **Privacy**:

"using personal data may improve public services by tailoring them based on personal characteristics or demographics, but compromise personal privacy because of high data demands."

Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Tension

- Using data to ***improve the quality and efficiency of services*** versus ***respecting the privacy and informational autonomy of individuals.***

Hypothetical illustration

- „A cash-strapped (*) public hospital gives a private company access to patient data (scans, behaviours, and medical history) in exchange for implementing a machine learning algorithm that vastly improves doctors' ability to diagnose dangerous conditions quickly and safely. „

(*) A Person or an organization that does not have enough money

- Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London*. Nuffield Foundation.

Hypothetical illustration

- „ The AI system will only be successful if the data is plentiful and transferable, which makes it hard to predict how the data will be used in advance, and hard to guarantee privacy and to ensure meaningful consent for patients. “

- Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London*. Nuffield Foundation.

Catalogue of Tensions

- **Privacy *versus* Transparency**:

“the need to respect privacy or intellectual property may make it difficult to provide fully satisfying information about an algorithm or the data on which it was trained.”

Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Catalogue of Tensions

- **Personalisation** *versus* **Solidarity**:

"increasing personalisation of services and information may bring economic and individual benefits, but risks creating or furthering divisions and undermining community solidarity. "

- Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Tension

- Reaping the benefits of increased personalisation in the digital sphere versus enhancing solidarity and citizenship.

Hypothetical illustration

„A company markets a new personalised insurance scheme, using an AI system trained on rich datasets that can differentiate between people in ways that are so fine-grained as to forecast effectively their future medical, educational, and care needs.

The company is thus able to offer fully individualised treatment, better suited to personal needs and preferences.“

- Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Hypothetical illustration

- “The success of this scheme leads to the weakening of publicly funded services because the advantaged individuals no longer see reasons to support the ones with greater needs. “

- Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Catalogue of Tensions

- **Convenience *versus* Dignity**:

“increasing automation and quantification could make lives more convenient, but risks undermining those unquantifiable values and skills that constitute human dignity and individuality. “

Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Tension

- Using automation to make people's lives more convenient versus promoting self-actualization and dignity.

Hypothetical illustration

- „An AI system makes possible an all-purpose automated personal assistant that can translate between languages, find the answer to any scientific question in moments, and produce artwork or literature for the users' pleasure, among other things."

Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Hypothetical illustration

- “Its users gain unprecedented access to the fruits of human civilization but they no longer need to acquire and refine these skills through regular practice and experimentation.
- These practices progressively become homogenized and ossified and their past diversity is now represented by a set menu of options ranked by convenience and popularity. “
- Source: Whittlestone, J et al

Catalog of Tensions

- **Satisfaction of preferences** *versus* **Equality**:

"automation and AI could invigorate industries and spearhead new technologies, but also exacerbate exclusion and poverty. "

Source: Whittlestone, J et al (

How to Identify further tensions

„Thinking about tensions could also be enhanced by systematically considering different *ways* that tensions are likely to arise. “

Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Identifying further tensions

- **Winners versus losers**. Tensions sometimes arise because the costs and benefits of ADA-based technologies are unequally distributed across different groups and communities.
- **Short term versus long term**. Tensions can arise because values or opportunities that can be enhanced by ADA-based technologies in the short term may compromise other values in the long term.
- **Local versus global**. Tensions may arise when applications that are defensible from a narrow or individualistic view produce negative externalities, exacerbating existing collective action problems or creating new ones.
- Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London*. Nuffield Foundation.

Resolving the tensions

- „The best approach to resolving a tension will depend on the nature of the tension in question. „

Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Use a Classification of Ethical Tensions

- **True Ethical dilemma**, i.e. "a conflict between two or more duties, obligations, or values, both of which an agent would ordinarily have reason to pursue but cannot";
- **Dilemma in practice**, i.e. "the tension exists not inherently, but due to current technological capabilities and constraints, including the time and resources available for finding a solution";
- **False dilemmas**, i.e. "situations where there exists a third set of options beyond having to choose between two important values".

Source: Whittlestone, J et al (2019)

Trade Offs

- „To the extent that we face a true ethical dilemma between two values, any solution will require making trade-offs between those values: choosing to prioritize one value at the expense of another. „

- Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation.

Trade Offs

- „ For example, if we determined that each of the tensions presented above could not be dissolved by practical means, we would need to consider trade-offs such as the following“:

Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation

Examples of Trade Offs

- Trade-off 1: Judging when it is acceptable to use an algorithm that performs worse for a specific subgroup, if that algorithm is more accurate on average across a population.

- Trade-off 2: Judging how much we should restrict personalization of advertising and public services for the sake of preserving ideals of citizenship and solidarity.
- Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation

Examples of Trade Offs

- Trade-off 3: Judging what risks to privacy it is acceptable to incur for the sake of better disease screening or greater public health.

- Trade-off 4: Judging what kinds of skills should always remain in human hands, and therefore where to reject innovative automation technologies.

- Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence:*

How Trade Offs Should be made

- The difficult question is how such trade-off judgments should be made.
- In business and economics, solutions to trade-offs are traditionally derived using **cost-benefit analysis (CBA**): where all the costs and benefits of a given policy are converted to units on the same scale (be it monetary or some other utility scale such as well-being) and a recommendation is made on the basis of whether the benefits outweigh the costs.

Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation

Cons of CBA

- „CBA does not recognize the fact that values are vague and unquantifiable and that numbers themselves can hide controversial value judgments, and finally, the very act of economic valuation of a good can change people’s attitude to it (this explains why applying CBA to environmental or other complex and public goods attracts so much controversy) „

- Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation

Consultation and inclusive public deliberation

- „ But one approach is that legitimacy of any emerging solution can be achieved through consultation and inclusive public deliberation“

- Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation

Trade Offs for Dilemmas in practice

In these situations we face a choice:

- **To put the technology to use in its current state**. In this case, we will need to determine and implement some legitimate trade-off that sacrifices one value for another. This will involve the same kind of work as described for true dilemmas.
- **To hold off implementing this technology** and instead invest in research on how to make it serve all the values we endorse equally and maximally.

Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), *London.* Nuffield Foundation

Useful Questions

- **Can the most accurate predictive algorithms be used in a way that respects fairness and equality?**

Where specific predictive algorithms are currently used (e.g. in healthcare, crime, employment), to what extent do they discriminate against or disadvantage specific minorities?

Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S.(2019), *London.*
Nuffield Foundation

Useful Questions

- **Can the benefits of personalization be reaped without undermining citizenship and solidarity?**

In what specific ways might different forms of personalization undermine these important ideals in future? How can this be addressed or prevented?

Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S.(2019), *London.*
Nuffield Foundation

Useful Questions

- **Can personal data be used to improve the quality and efficiency of public services without compromising informational autonomy?**

To what extent do current methods allow the use of personal data in aggregate for overall social benefits, while protecting the privacy of individuals' data?

Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S.(2019), *London.*
Nuffield Foundation

Useful Questions

- **Can automation make lives more convenient without threatening self-actualisation?**

Can we draw a clear line between contexts where automation will be beneficial or minimally harmful and tasks or abilities should not be automated?

Source: Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S.(2019), *London.*
Nuffield Foundation

Reference

Whittlestone, J et al (2019) – *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S.(2019), *London.*

Nuffield Foundation