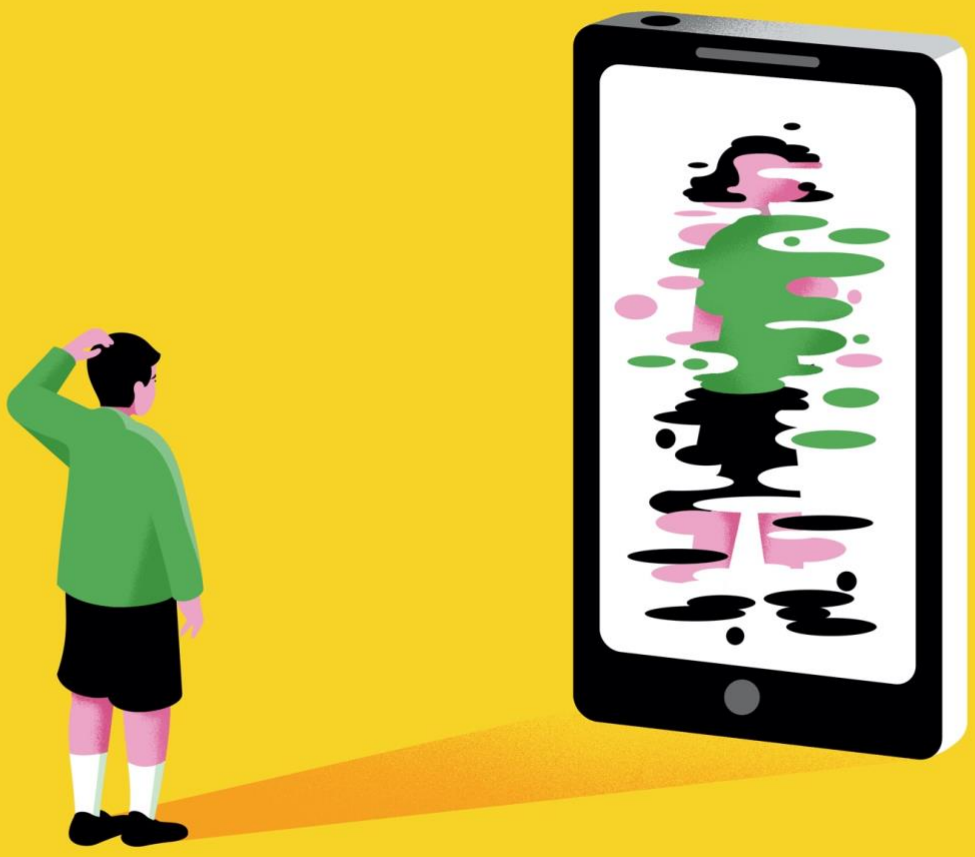
June 2022

Shedding light on AI

A framework for algorithmic oversight



5RIGHTS FOUNDATION

Table of Contents

Introduction

There is increasing global consensus that digital services that impact on children must be designed for their use. Childhood is a time of experimentation and personal growth, and while no environment is entirely risk free, digital services "likely to be accessed by children" must be designed to be private, safe and rights respecting - by default.

Central to the digital world is artificial intelligence, commonly referred to as AI. AI is not a standalone or fixed technology but plays a part in automated decision-making (ADM) systems and many other data-driven features common across digital services. Automated systems shape the experiences of children and young people in the digital world, both as a result of their direct engagement, for example receiving friend/follower or content recommendations, and from systems that they may not interact with directly, for example automated decision-making used to allocate welfare funding.

Much emphasis is put on the challenges of regulating new and emerging technologies, but AI is not new. The term 'AI' was coined in the 1950s to describe the science and engineering of machines making automated choices against specific criteria based on available information. In many ways, the word 'intelligence' is used to give humans confidence in the efficacy and authority of machine-made choices. Since then, huge advancements in the application of AI and greater availability of data have led to more sophisticated, data-driven decision making. Systems that use AI are still human-made with specific objectives, design goals, chosen inputs, a set of rules by which information is given importance or weight, and a combination of outcomes and outputs. At each of these stages, automated decisions are made that are often imperceptible to those they impact, particularly if they are a child.

Automated decision-making sits behind features that are ubiquitous across services likely to be accessed by children. It can support children to navigate the online world and the mass of content available, and help them to identify activities and outcomes that are useful or beneficial to them. But there are also many situations when automated decision-making systems undermine their rights or put them at risk. For example;

- **Friend recommendations** are made by 75% of the most popular social networking sites.[1] These connect users based on the data profiles the platform has built about them, irrespective of their age, which has been found to enable predators to contact children.[2]
- **Misinformation** is spread and amplified by automated systems that frequently promote content that is most likely to engage users, irrespective of its quality or impact. In 2020, vaccine misinformation alone was worth up to $1bn for the largest services.[3]
- **Nudges** are optimised to encourage children to make in-app purchases or engage with gambling-style features. In 2019, British children collectively spent £270 million on loot boxes and other in-app purchases.[4]

---

[1] But how do they know it is a child?, 5Rights Foundation, 2021.

[2] Instagram sends predators to accounts of children as young as 11, The Times, 2019.

[3] The Anti-Vaxx Industry: How Big Tech powers and profits from vaccine misinformation, Center for Countering Digital Hate, 2021.

[4] Young People Losing Millions to Addictive Gaming – REPORT, Safer Online Gambling Group, August 2019.

- **Harmful material** that violates the community rules of services is recommended to children, including material promoting self-harm or suicide behaviours, disordered eating and pornography.[5]

New online safety legislation in Europe and legislative proposals across the world, such as the UK, Canada and Australia offer a vision of what a responsible digital world looks like. However, to meet the objectives of online safety legislation in both spirit and letter, regulatory authorities must not only have the tools but a duty to investigate algorithms on behalf of children, and an agreed standard by which to assess them. A duty of this kind with the four-step process described in this paper would ensure the risks to children created by algorithms are identified, eliminated, mitigated or effectively managed.

This four-step process is platform neutral and can be applied across different sectors, including but not limited to social media, entertainment, health and education. It can also be applied to different parts or features of a service, including advertising, content recommendation, moderation and reporting.

Such a regime would give clarity to businesses in fulfilling their safety duty to children and power to the regulator to inquire, analyse and assess whether a system is conforming to requisite standards. When new risks and harm are revealed, they can act as an early warning, especially when that harm was an unintentional by-product of an automated decision-making process optimised for another purpose.

This short paper builds on the work of many in the international community, notably the Centre for Data Ethics and Innovation[6], UNICEF[7], IEEE[8], the Ada Lovelace Institute9 and the Council of Europe[10]. We are grateful for their expertise and recognise that this practical application of their work could not have been done without their thoughtful and detailed insights.

Thanks are due also to Dr Rebekah Tromble, Associate Professor in the School of Media and Public Affairs and Director of the Institute for Data, Democracy, and Politics at George Washington University. Dr Tromble developed the four-step model articulated in this report.

5Rights is committed to building the digital world young people deserve. That world is one in which they share the benefits of digital engagement as participants, citizens and consumers, and in which businesses respect and uphold their existing rights and respond to their needs and evolving capacities - automatically.

---

[5] Pathways: How digital design puts children at risk. 5Rights Foundation, 2021

[6] In November 2020, the Centre for Data Ethics and Innovation conducted a review into bias in algorithmic decision making and made recommendations to the government and regulators designed to produce a step change in the behaviour of organisations making life changing decisions on the basis of data.

[7] UNICEF's Draft Policy Guidance on AI for Children is designed to promote children's rights in government and private sector AI policies and practices, and to raise awareness of how AI systems can uphold or undermine these rights. The policy guidance explores AI and AI systems and considers the ways in which they impact children. It draws upon the Convention on the Rights of the Child to present foundations for AI that upholds the rights of children.

[8] The IEEE (Institute of Electrical and Electronics Engineers) has a global Initiative on the ethics of autonomous and intelligent systems. Its aim is to move from principles to practice with standards projects, certification programs, and global consensus building to inspire the ethically aligned design of autonomous and intelligent technologies.

[9] Ada Lovelace Institute are developing tools to enable accountability of public administration algorithmic decision-making, such as a typology and a public register.

[10] Council of Europe and Artificial Intelligence. Council of Europe, April 2022.

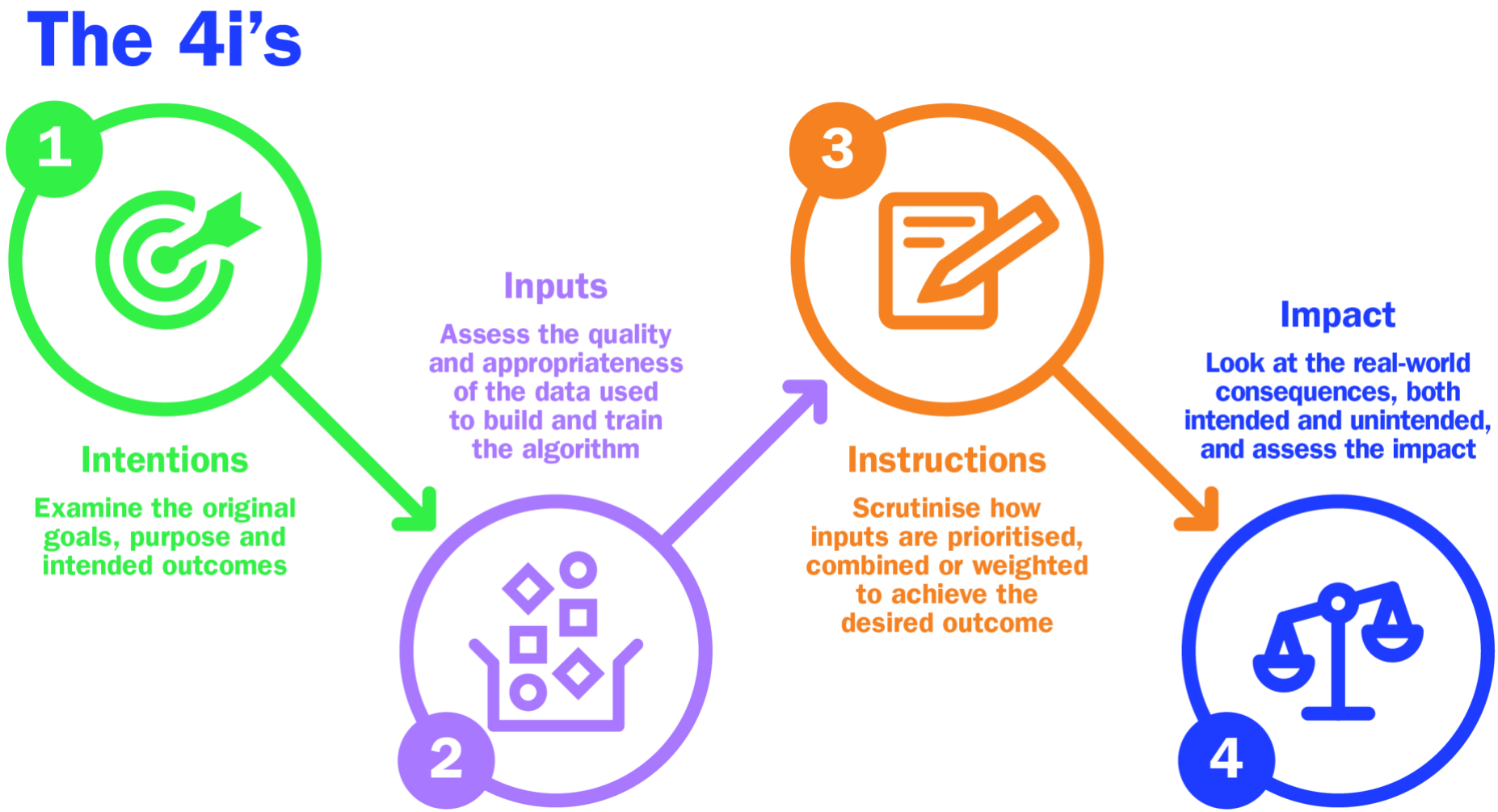
Definitions

- **Artificial intelligence** (AI) describes when machines are able to mimic the problem-solving and decision-making capabilities of the human mind.
- **Machine learning** is a branch of artificial intelligence (AI) that employs computational algorithms to detect patterns in—and learn iteratively from—data, generating output with minimal human intervention and improving over time.
- An **algorithm** is a sequence of instructions or set of rules designed to complete a task or solve a problem.
- **Automated decision-making** is the process of making a decision by automated means, without any human involvement. These decisions can be based on factual data, as well as on digitally created profiles or inferred data.[11]
- **Algorithmic bias** is commonly used to describe an automated system that produces results that discriminate against or disadvantages groups of people (for example, based on age, disability, gender, or race).
- **Algorithmic fairness** is an automated system that produces results that do not discriminate against nor systematically disadvantage groups of people: it also seeks to ensure that automated systems do not violate rights, exploit vulnerability, manipulate, nor withhold information in a way that impairs one's ability to make informed choices.
- A **child** is a person under the age of 18.[12]
- **Document analysis** is a research method involving the review and interpretation of written materials such as emails, legal records, and meeting notes, designed to gather evidence on the topic being studied and answer specific questions.
- **Code analysis** is a way of assessing how an algorithm is structured and how it might function in practice without actually executing the program, allowing errors or vulnerabilities to be detected.
- **Variables** are individual items in a dataset being analysed, for example age, gender and location.

[11] What is automated individual decision-making and profiling? Information Commissions Office, 2018.

[12] Article 1 of the United Nations Convention on the Rights of the Child states "a child means every human being."

Four step model for regulating AI



The four steps of AI oversight, also known as the 4 Is.

The four-step model set out below offers a mixed method approach to algorithmic oversight. It describes how regulators can evaluate each element of an automated decision-making process, from the goals, inputs, implementation and outcomes, to ensure that applications of AI meet the established rights and needs of children, as set out in:

- The United Nations Convention on the Rights of the Child to which the UK is a signatory. General Comment 25 on children's rights in relation to the digital environment is the authoritative document which sets out the relevance of the convention to the digital world.[13]
- General Data Protection Regulation (GDPR)[14] and the Age Appropriate Design Code,[15] and the associated guidance from the Information Commissioners Office, which set out the standards providers of digital products and services must meet in relation to children's data.
- Existing laws and protections that pertain to children, such as the Equality Act 2010, Children Act 1989, and some consumer laws, as well as government guidance, which may also be relevant when considering the impact of AI systems on children.[16]

Any assessment of automated decision-making systems must have the flexibility to uncover harms that are currently unknown or not anticipated. It must also allow for potential improvements or benefits to be identified so that they might be shared and used to guide best practice.

1. Understand the design goals

Aim:

Algorithms are formulated with a purpose and intended outcomes. In assessing the fairness and appropriateness of algorithms, it is important for the regulator to understand the original intent and goals of its creators and how those goals evolved over time, by asking the following questions:

a. What was/were the problem(s) or challenge(s) those designing the algorithm set out to address?

b. What was/were the intended outcome(s)?

c. Why was this product, feature or process considered necessary?

d. Who was involved in defining the problem(s) and desired outcome(s)—including internal and external stakeholders? What was their role in shaping the understanding of the problem(s) and desired outcome(s)?

e. How and why did any of these things change over time?

Method:

- Undertake interviews with stakeholders

[13] General Comment No. 25 (2021) on children's rights in relation to the digital environment, UNCRC, 2021

[14] General Data Protection Regulation. Official Journal of the European Union, 2016.

[15] Age Appropriate Design Code. Information Commissioner's Office, September 2020.

[16] For example, advice from the Chief Medical Officer or guidance from the Department of Education.

- Analyse information - product development documents and internal communications such as emails and meeting notes in which the algorithmic product was discussed.

2. Consider the data inputs

Aim:

Every algorithm contains a series of inputs — data points and variables that can be thought of as the "ingredients" of the algorithm. Unfair, discriminatory or biased outcomes are often the result of problematic data ("garbage in, garbage out"). It is therefore essential that any framework intended to examine algorithmic fairness assess the quality and appropriateness of the data used to build and train the algorithm, by asking the following:

a. What features (variables) did the algorithm's designers want to include as inputs and why?

b. Were they able to include those features? Did they have to settle for proxies and/or exclude some features altogether and why?

c. What dataset(s) was/were used as input(s) for building, training, and testing the algorithm?

d. Were other datasets considered? For training/testing? For final implementation? If not, why not?

e. If so, what were the perceived advantages and disadvantages, strengths and weaknesses of this/these datasets compared to other options?

f. Were multiple datasets and/or features tested? If so, how were they evaluated? And why were the final datasets/features selected?

g. Who had input into these decisions, and what was their role in the process?

Method:

- Undertake interviews with stakeholders
- Analyse information - product development documents and internal communications
- Code analysis
- Data sample analysis.

3. Assess the model selection and execution

Aim:

If data inputs are the "ingredients" of an algorithm, the mathematical model and its parameters offer the instructions for how to put the algorithmic recipe together. They lay out how the inputs should be combined, at what point and in what amount, as well as the ways in which those inputs might be altered or transformed. Careful scrutiny of the model and the assumptions it is built upon is needed to assess its appropriateness. Note that such scrutiny is possible even with machine learning algorithms. The questions to consider as part of this scrutiny include:

a. What is the mathematical formula/model applied?

b. Why was this model selected?

c. What assumptions are built into this model?

d. Did those designing or implementing the algorithm deviate from any of the assumptions built into the model? If so, how and why?

e. Within the model, what is being optimised? How is this optimisation carried out (e.g., how are the various features weighted)?

f. How and when was the model tested and changed/updated?

g. When changes were made, what were the reasons for making those changes?

Method:

- Undertake interviews with stakeholders
- Analyse information - product development documents and internal communications
- Code analysis
- Implementation experiments (e.g., running independent tests on real or synthetic data, including on platform).

4. Identify outputs and outcomes

Aim:

After an algorithm is launched, it will generate certain outputs. It is important to examine these outputs to reveal whether the model performs as intended. However, at this stage, it is also important to look at the actual outcomes — the real world impacts generated by the algorithm(s) and its uses.

The previous three steps help to determine why and how something went wrong, what elements of the design and implementation result in discrimination, disadvantage, exploitation, manipulation, or rights violations. However, the output (step four) is likely to be the first place that harm is identified and the stage at which it is shown that the four step process is necessary.

Many observers note that algorithms are not autonomous, neutral entities. They are designed by people, with all the biases, blind spots, and other foibles associated with being human. It is therefore crucial to examine the interplay of technical features on the one hand, business decisions and human interactions on the other. The regulator will need researchers and investigators with training in the social sciences as well as computer scientists to conduct such assessments.

Below we lay out the three lenses through which to examine algorithmic outputs and outcomes. First, we describe assessments of the ways in which relevant companies interpret outputs and outcomes, as well as their techniques for mitigating perceived harms. Second, we outline a broad approach for considering how users interact with and are impacted by algorithms. Finally, we discuss broad approaches to uncovering impacts on society as a whole.

Companies

Aim:

To examine how either the company that designed the algorithm or companies that make use of those algorithms evaluate outputs and outcomes.

Questions:

a. What model outputs (variables) does a company use internally? (I.e., What outputs matter to them and why?) In what ways do they use these outputs?
b. What is the internal process for evaluating the performance of an algorithm? What standards are applied? What metrics are applied? By whom?
c. What is the internal process for determining whether an algorithm should be changed? Who is involved in this process? Who makes final decisions and how?
d. What, if anything, is the company doing to assess larger impacts on users and society?
e. If such assessments occur, are they ad hoc or systematic?
f. What techniques and methodologies are used for such an assessment? What standards and metrics are applied? Who is involved in this process and how?
g. Are changes ever made to algorithms on the basis of such assessments? What is the process for doing so? Who is involved in this process? Who makes final decisions and how?

Method:

- Interviews
- Document analysis
- Code analysis.

Users

Aim:

To assess whether users' reasonable expectations for how they interact with and what they expect from an algorithm align with the actual outcomes, and whether any harms (either perceived by the user or not) accrue.

Questions:

a. What, if anything, do users understand the algorithm to be doing? Are they even aware that an algorithm is involved? If they are, do they perceive specific advantages and disadvantages to the algorithm?

b. What do users expect from the algorithm? Are outcomes aligned with those expectations?

c. Is the algorithm creating disparities between users and non-users and/or between different types of users?

d. Is the algorithm limiting user choice(s)? If so, in what ways? And what are the consequences (positive or negative) of those limitations?

e. Does the algorithm directly or indirectly exploit user vulnerabilities?

f. Does it directly or indirectly manipulate users?

g. Does it violate users' rights or contribute in any way to the violation of those rights?

Method:

- User surveys and interviews
- (Controlled) experimental user studies.

Societal impacts

Aim:

To understand the social, financial, environmental and human impacts of automated decision-making systems.

Questions:

a. Is the algorithm contributing directly or indirectly to social harms? If so, in what ways? And to whom? Is the harm caused by certain features of the algorithm? Can these harms be mitigated by changes to the algorithm? Can these harms be mitigated without causing harm to others?

b. Is the algorithm benefitting certain members of society? If so, are those benefits accrued fairly and equitably?

c. Is the algorithm benefitting society as a whole? If so, in what ways? Can those benefits be amplified or expanded?

d. Are there “best practice” lessons to be learned from the design and implementation of this algorithm?

Method:

- A variety of social, scientific and humanistic research designs.

Regulatory duty to investigate

Children cannot be expected to understand or take action against automated decision-making or algorithmic unfairness, it is unlikely that they have the developmental capacity, the knowledge or the resource to understand the subtle, cumulative or even acute nudges and impacts those automated systems have on their online experience. In fact – many children do not understand that an algorithm could be responsible for introducing them to a 'suggested friend' nor do they have the tools to prevent an onslaught of automated harmful material. Regulators must be given not only the powers to interrogate automated systems but create the expectation that they will be actively analysing automated decision-making systems and algorithms of services that impact on children – a duty to investigate.

The proposed duty is similar to that of the UK's Financial Conduct Authority, which has a duty to investigate when it appears that a regulated person or investment scheme has failed to protect consumers or might have a significant adverse effect on the financial system or on competition. Similar duties are proposed for a new pro-competition regime that would give the Digital Markets Unit (part of the Competition and Markets Authority) a duty to 'monitor' markets and the activities of firms to identify breaches of the statutory code of conduct.

In order to fulfil this duty, regulators must have the expertise, resource, and processes in place to scrutinise the design goals, data inputs, model selection and outputs and outcomes of algorithms. Where there is evidence to show such systems are discriminating against or systematically disadvantaging children or violating their rights, regulators should set out a mandatory course of action for compliance.

While transparency is a key component of the four-step process set out above, decades of research show transparency alone can result in layers of obfuscation and does not always result in better systems or more positive outcomes. The value of transparency lies not in the availability of information itself, but in the way it allows for scrutiny and accountability. A duty for regulators to undertake the four steps on automated decision-making systems that impact on children would deliver that accountability.

Companies often use commercial sensitivity as a defence to usurp transparency reporting requirements. On the whole, this should be resisted, and where there are legitimate commercial sensitivities, the regulator must have the power to maintain private oversight.

Conclusion

A child must not be asked to police the automated decisions of the tech sector. The industry is worth over $5 trillion to the world economy[17] and is central to children's lives and life outcomes. Algorithmic oversight is critical if the next generation of digital technologies, products and services are to offer children safety and respect for their rights, by design.

The four-step model of algorithmic oversight will allow meaningful oversight of the goals, inputs, implementation and outcomes of algorithms and automated decision-making systems. This transparency will drive a change in corporate behaviour that meets the expectations of parents and children and fulfil the promises government has made to them. By giving the regulator a duty to interrogate automated decision-making systems on behalf of children, and service providers a clear process by which it will be done, the risks to children from automated decision-making systems can be reduced – by default and design.

There is no silver bullet to fix all the ills of the digital world or to guarantee children will be safe from harm, either through regulation or technological development. But the argument that regulation and accountability stifle innovation or impose limits on a child's freedom in the digital world is simply untrue. Each wave of regulation has been met with creative and practical solutions. As lawmakers across the globe look to the UK to continue its singular place as leader in online safety for children algorithmic oversight would continue that march.

It is in the interests of all parties to have a more equitable and trustworthy system of oversight that allows growth and innovation but which reduces negative outcomes for children.

To do nothing is no longer an option.

[17] Distribution of the information technology (IT) industry worldwide from 2019 to 2022, by region, Statista, 2022.