

CHILDREN & AI DESIGN **CODE**

A protocol for the development
and use of AI systems that impact children

Preface

The eye-catching announcement in November 2022 of OpenAI's ChatGPT put the issue of AI onto the front pages of newspapers and into the public's imagination. Since then, we have seen almost constant announcements that AI is the answer to many of the world's intractable problems – from cancer to declining productivity and delivery of public services, and maybe even mortality. At the same time, those inside and out of the AI sector have issued stark warnings of its power, of the unfettered race to control it, and that we are on a path to untruth, unemployment, antonymous weaponry, and possibly even human annihilation.

Both narratives have been running for many more decades than they have had our attention. AI is not new. It has been in development for more than 70 years (see 'A brief history of AI'), and both creative and factual writers have been warning of its overwhelming power. But these narratives have obscured the more quotidian issues of AI, a powerful technology being developed for the most part by private businesses with narrow commercial or ideological goals, its propensity for illusion, and the fact that it is a numbers game – giving the most likely statistical response to a question rather than evidenced truth. Perhaps most importantly of all, that it has been developed in the wake of a tech sector that has negotiated free reign for its products and services, developed with little oversight, no liability, and few responsibilities.

These three privileges have allowed the sector to grow at an enormous rate, while the discord and harms that have come in their wake have been characterised as collateral damage to the greater good of technological progress. It is increasingly clear that the benefits of the technology have also accrued unevenly, and to some, dangerously, as a handful of private individuals own both the means of production and distribution to consolidate power, money, and ideological dominance.

Over the last ten years, some jurisdictions have pushed back, particularly in relation to children. The introduction of regulatory or legal protections such as the UK's Age Appropriate Design Code,¹ the EU's Digital Services Act² and AI Act,³ Ireland's Online Safety and Media Regulation Act,⁴ Singapore's Online Safety (Miscellaneous Amendments) Act⁵ and accompanying Code of Practice for Online Safety,⁶ and Australia's Online Safety Act⁷ have led to significant redesign of products, and in some cases made companies legally accountable for their actions.

The Children and AI Design Code builds on those initiatives and on the understanding that we all have a shared responsibility for children, including those who build and benefit from technology. AI may be a wonder and generative AI a previously unimaginable breakthrough, but if AI, like other tech before it, moves fast and breaks things, we must, at a minimum, act on the consensus that we may not allow it to break our children.

The Code is practical and actionable and sits on the shoulders of many organisations, computer scientists, academics, and experts who have generously given their time and expertise to its creation. The Code will not irradiate all risk, and neither will it inhibit development in any meaningful way – other than to make the design of AI systems conscious in order to prevent foreseeable harm to children.

Decisions on how and on what basis AI is adopted into public and private life must consider children's needs from the outset, and by design. The proposed Children and AI Design Code does just that.

Thanks are due to the enormous number of contributors to this project. Many (but not all, for reasons of anonymity) are listed below. Thanks also to Alexandra Evans who took charge of much of the drafting, repeatedly responded to input, and navigated towards a consensus across a large number of experts.

As ever, the greatest thanks go to the children and young people who gave their opinions on uses of AI. The pages that follow codify their desire for a fair, exciting, and child-conscious digital world.



BARONESS BEEBAN KIDRON
5Rights Foundation Founder & Chair

MARCH 2025

Contributors

Dr Ayça Atabey, University of Edinburgh and Digital Futures for Children centre

Nicholas Dunn, *student*

Dr Pedro Hartung, *CEO*, Alana Foundation

Professor Ali Hessami, *Technical Editor and Chair* of the IEEE P7000:2021 Standard on Addressing Ethical Concerns in System Design

Louise Hooper, *public law and human rights barrister*, Garden Court Chambers

Anja Kaspersen, *Director* at IEEE SA

Professor Sonia Livingstone, *Director of Digital Futures for Children centre*, London School of Economics and Political Science

Dr Manolis Mavrikis, *Professor of Artificial Intelligence in Education*, UCL Knowledge Lab, University College London

Moira Patterson, *Director* at IEEE SA

William Perrin, OBE

Dr Andrew Serazin, *Senior Research Fellow*, Reuben College and Director of the Global Challenges Programme

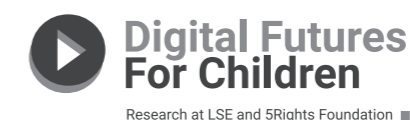
Vaishnavi J, *Founder*, Vyanams Strategies (Vys) / Nell Watson, President, EURAIO

Steve Wood, *Founder*, PrivacyX Consulting and former UK Deputy Information Commissioner

Dr Jun Zhao, Department of Computer Science, University of Oxford

5Rights Youth Ambassadors

And others who prefer to remain anonymous.



PART 1	
Context	8
1.1 Background	8
1.2 Overview of the Code	9
1.3 What is an AI system?	9
1.4 Who is the Code for?	12
1.5 Compatibility with existing frameworks and processes	12
1.6 When do AI systems impact children?	14
PART 2	
Key Considerations	15
2.1 Supply chain	15
2.2 AI lifecycle	16
2.3 Context	17
2.4 Testing and metrics	17
2.5 Stakeholder engagement	18
2.6 Children's rights and capacities	18
2.7 Diversity and inclusion	20
2.8 Proportionality	20
2.9 Role of parents or carers	20
PART 3	
Criteria	21
3.1 Developmentally appropriate	21
3.2 Lawful	21
3.3 Safe	21
3.4 Fair	21
3.5 Reliable	21
3.6 Provide redress	22
3.7 Transparent	22
3.8 Accountable	22
3.9 Uphold children's rights	22

PART 4	
Common risks to children from AI systems	23
4.1 Unfairness	23
4.2 Harmful content and activity	23
4.3 Privacy	24
4.4 Security	24
4.5 Capture	24

PART 5	
The Code	25
5.1 Preparation	27
5.2 Intentions	30
5.3 Data	31
5.4 Development	35
5.5 Deployment	38
5.6 Monitoring	40
5.7 Transparency	41
5.8 User reports and redress	44
5.9 Retiring and moving on	47

PART 6	
Further context and definitions	51
6.1 Childhood development	51
6.2 Team member roles and responsibilities	55
6.3 Definitions	62
6.4 Children and AI Design Code requirements checklist	68
6.5 Snapshot case studies	74
Endnotes	92

PART 1

1.1 Background

The transformative effects of the widescale use of AI systems will shape almost all aspects of society. Governments and international institutions are now grappling with how to establish standards and oversight to ensure these technologies are developed in the interests of their populations and humanity. There are many, including some of those who are responsible for creating the new models, who believe that policy and law makers are moving too slowly.⁸

Children make up 30% of the global population and are disproportionately early adopters of technology, including products and services that use or embed AI.⁹ Yet too often their needs, rights, and views are not represented in the public and policy debate on AI.¹⁰ So while much has been said or suggested about oversight for AI, little has focused on children, and the practical measures needed to ensure that children's rights and development needs are met.

While world leaders consider, and fail to agree, treaties and legislation that would ensure the safe and equitable development of AI, children are growing up in a world that is increasingly shaped and judged by AI. The Code takes a pragmatic and practical view. It sets out a process to identify, evaluate, and mitigate the known risks of AI to children and prepare for the known unknowns. It requires those who build and deploy AI systems to consider the foreseeable risks to children by design and default.¹¹

In 2023, the Alan Turing Institute published *AI, children's rights, & wellbeing*,¹² a review of key transnational frameworks that are relevant to children and AI. It identified a regulatory gap and called for a practical application of its findings. The Code builds on this work by operationalising these frameworks into an implementable process.

The Code demands engagement from companies that create, deploy, or use AI. Its questions demand answers that if answered directly and in good faith, will describe a path to conscious, rights-respecting innovation.

Much of the work described requires a team. It may be that existing teams can be reorganised or new ones created, but the team must meet the needs of the Code. The scale and purpose of AI products and services are infinite, so the Code provides a set of actions that must be demonstrably undertaken by people with the correct skills and authority to act. In this iteration the Code is voluntary for those who proactively want to consider children, but as AI becomes more central to all our lives, the Code provides the bar for regulatory initiatives that seek to support children across the globe.

No one can claim to know the totality of benefits and harms AI systems will bring. Even among the most embedded experts (including Nobel Prize winners) there are differences of opinion. The Code offers a continuous process that can be used at any stage of the lifecycle of an AI system to ask the correct questions. This will ensure conscious technical development and support innovation that impacts positively on children and the world they inhabit. In doing so it is one step to building the digital world children deserve.

1.2 Overview of the Code

PART ONE provides the context to the Code.

PART TWO provides advice and guidance on key considerations that are relevant at all stages of the Code.

PART THREE sets out the criteria that an AI system that impacts children must meet.

PART FOUR describes potential risks to children.

PART FIVE is the Code itself. It includes a checklist of key actions and guidance at each stage of the lifecycle of an AI system.

PART SIX provides further information, including key definitions and concepts, as well as stages of child and adolescent development and snapshot case studies to illustrate how the criteria might apply.

The Code has been developed with input from experts from a wide range of disciplines and fields. Future iterations will be needed to reflect further contributions, corrections, and debate, as well as the rapid evolution in AI capabilities.

1.3 What is an AI system?

Although there is no universally agreed definition of an AI system, there is a high level of consensus.

In the European Union's (EU) Artificial Intelligence (AI) Act, an 'AI system' means 'a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.'¹³

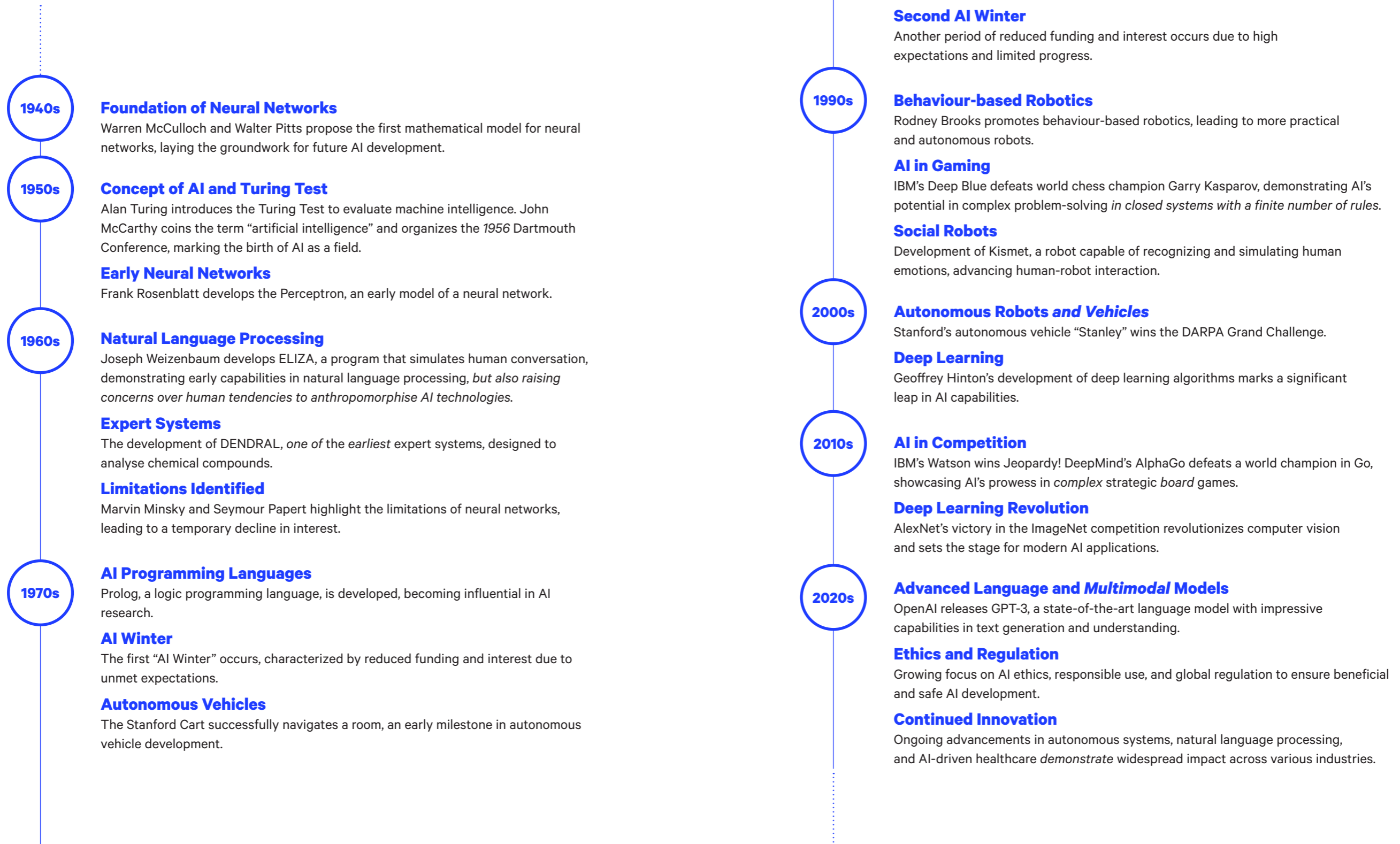
The Organisation for Economic Co-operation and Development (OECD) defines an AI system as 'a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.'¹⁴

In the US, the National Institute of Standards and Technology's (NIST) Artificial Intelligence Risk Management Framework (AI RMF 1.0) refers to an AI system as 'an engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy (Adapted from: OECD Recommendation on AI:2019; ISO/IEC 22989:2022).'¹⁵

A brief history¹⁶ of AI Fig1.

This timeline highlights key themes and milestones in the evolution of AI from theoretical concepts to practical applications and its growing influence on various aspects of society.

It was generated by ChatGPT with additions by a human being (in italics).



1.4 Who is the Code for?

The Code applies to AI systems of all kinds. It is applicable to public and private sectors, small and large companies, and, as seen from the case studies, across domains – education, health, media etc.

The Code is primarily for those designing, adapting, or deploying an AI system that impacts children, but it can also be used by:¹⁷

- **Process assessors**, to serve as a reference model for third parties to assess conformity of standards for AI systems that impact children.
- **Organisations** wishing to understand what aspects of AI technologies impact children.
- **Children are likely to engage directly or indirectly with an AI system.** For example, a child using a search engine or a surveillance system used to monitor spaces where children are present.
- **An acquirer or supplier**, to guide acquirers of AI systems or component parts (e.g. data sets or foundation models) to create due diligence processes or structure requests for assurances of minimum standards for AI systems that impact children.
- **Governments and oversight bodies**, to inform the work of policy makers, regulators, and standards bodies as they develop minimum standards for AI systems that impact children.

1.5 Compatibility with existing frameworks and processes

The Code reflects existing legislation and regulatory frameworks. As such, it is grounded in the emerging consensus on general principles of AI governance. The provisions of the Code apply to all children in all geographies of all ages. A 'child' is a person under the age of 18.¹⁸

The Code is comprehensive but can be incorporated into other design and governance standards, mechanisms, and practices, for example risk analysis, impact assessment, or mandatory requirements.

Whether used alone or in conjunction with existing practices, the Code is not a pick-and-mix but requires each action to be considered and enacted in full.

The Code is compatible with:

- United Nations' (UN) Convention on the Rights of the Child (UNCRC)¹⁹ and General comment No. 25²⁰ on children's rights in relation to the digital environment;
- European Union's (EU) AI (Artificial Intelligence) Act;²¹
- United States' Blueprint for an AI Bill of Rights²² and now repealed Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence;²³

- Council of Europe's Framework Convention on artificial intelligence and human rights, democracy and the rule of law²⁴ and Methodology for the risk and impact of artificial intelligence systems from the point of view of human rights, democracy and the rule of law (Huderia methodology);²⁵
- National Institute of Standards and Technology's (NIST) Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile.²⁶

It is informed by:

- Council of Europe's Guidelines to respect, protect and fulfil the rights of the child in the digital environment;²⁷
- International Telecommunication Union's (ITU) 2020 Guidelines on child online protection;²⁸
- UNESCO's Recommendation on the ethics of artificial intelligence;²⁹
- UNICEF and Ministry for Foreign Affairs of Finland's Policy guidance on AI for children;³⁰
- Office of the High Commissioner for Human Rights' (OHCHR) Artificial intelligence and privacy, and children's privacy – Report of the Special Rapporteur on the right to privacy;³¹
- Organisation for Economic Co-operation and Development's (OECD) Recommendation of the Council on children in the digital environment³² and its Companion document;³³
- World Economic Forum's Artificial intelligence for children;³⁴
- UN AI Advisory Body's Governing AI for humanity;³⁵
- Hiroshima Process International guiding principles for organizations developing advanced AI systems;³⁶
- Seoul Declaration for safe, innovative and inclusive AI;³⁷
- Global Digital Compact (2024);³⁸
- OECD's Recommendation of the Council on artificial intelligence;³⁹
- UN's draft resolution on artificial intelligence;⁴⁰
- Council of Europe's Mapping study on the rights of the child and artificial intelligence;⁴¹
- Children's Rights Impact Assessment, Digital Futures Commission;⁴²
- African Union's Continental Artificial Intelligence Strategy: Harnessing AI for Africa's Development and Prosperity;⁴³
- Paris AI Summit's Statement on Inclusive and Sustainable Artificial Intelligence for People and the Planet.⁴⁴

1.6 When do AI systems impact children?

An AI system impacts children if, across its lifecycle or supply chain:

- (a) Children's data forms part of the data set on which the AI system has been trained.** For example, a large multimodal model (LMM) that enables users to generate photo-realistic images that are trained on images of children.
- (b) Children's experience of a service or product is shaped by the AI system.** For example, a digital service that uses an AI system to determine when to send notifications to users that shapes what times of day children open the service, how often they do so, and how long they spend on it.
- (c) Children are likely to engage directly or indirectly with an AI system.** For example, a child using a search engine or moving through a space that is being monitored by a surveillance system.
- (d) An AI system generates outputs or outcomes that are likely to impact children.** For example, healthcare software that determines which groups are high risk for a certain infectious disease and must be included in a nationwide vaccination scheme.
- (e) The AI system influences decisions made by adults that impact children.** For example, an education assessment tool that uses AI to predict children's academic potential based on a standardised test or behavioural monitoring.

If there is uncertainty about whether the system is likely to impact children, consider what evidence is already available about the context and likely use cases for the AI system. If you conclude that there is not likely to be an impact, you must record this, and also whether this decision will be periodically reviewed.

The decision must, in any event, be reviewed if there is any change to the intended use.

If your AI system does impact children, you must follow all the requirements as set out in the Code.

PART 2

Key considerations

Part Two sets out the overarching considerations that are relevant to all aspects of the Code. It is intended to help you understand and meet the criteria. While you may need to consider some issues once, you must consider others throughout the lifecycle and deployment of an AI system.

2.1 Supply chain

AI system supply chains are relevant to all aspects of conformity with the Code. If the data sets or models that you are using do not conform, your AI system is also unlikely to conform, as are onward or further uses based on your AI system.

In this regard the Code mirrors the European Union's (EU) AI Act (Article 25)⁴⁵ in determining that if you supply your AI system to others, the contract must be clear on any restrictions or parameters of use as well as your continued responsibilities, for example in providing any information and assistance needed to operate the system in accordance with this Code. This constitutes current best practice.

2.1.1 Mapping the supply chain

You must ensure sufficient visibility over your data supply to confidently map both upstream and onward uses. You must be able to describe and document the due diligence steps you have taken to assure yourself that others have complied with the Code, and you must have a strategy in place for responding to emerging concerns or emergency events (e.g., a data contamination incident).

2.1.2 Upstream (data and systems)

AI systems are often built using data sets or models sourced from third parties, and new products and services frequently add more than one data source to that of existing AI models. A complex and/or opaque supply chain makes conformity less straightforward. It also amplifies the risk to children (e.g., exacerbating risks to privacy, reliability, accountability, redress, and fairness). The ability of the supplier to provide assurances of conformity with the Code will be a determining factor in what models or data sets you incorporate.

The IEEE's Draft standard for the procurement of artificial intelligence and automated decision systems (IEEE P3119)⁴⁶ establishes a uniform set of definitions and a process model for the procurement of AI and automated decision systems (ADS). Government entities can use this to address sociotechnical

and responsible innovation considerations to serve the public interest. While its focus is on government procurement, it acts as a useful bar for all procurement processes.

2.1.3 Future use by others in the supply chain (onwards supply)

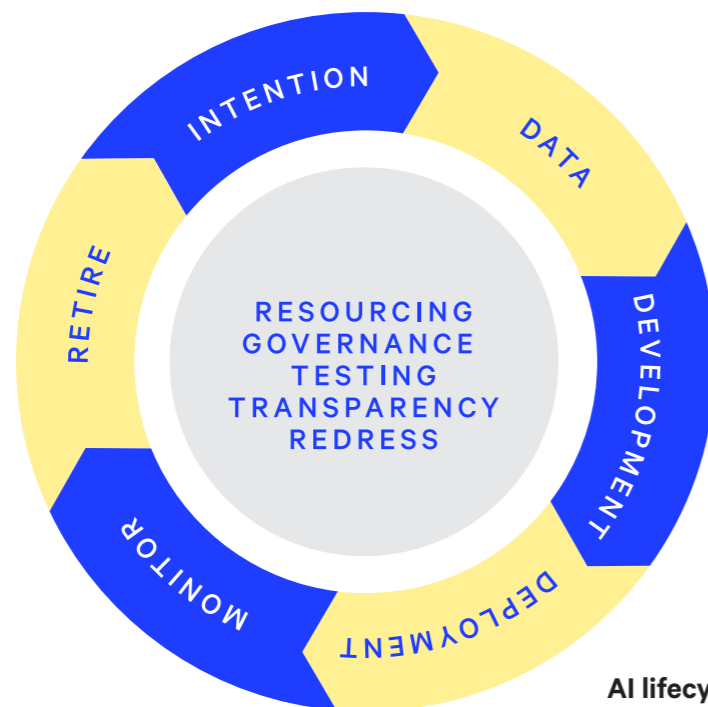
Negative impacts of your AI system may be perpetuated through further use of your AI system by other organisations. For example, if your AI system that recommends news articles is found to have flaws or inaccurate outcomes, other organisations that use your system will further amplify the spread of inaccurate news. You must have systems in place to mitigate the onward impacts of your system, and where possible, your contracts and terms of use should require onward use to be Code-complaint. There is further detail about the uses of data in the Code itself (Section 5.3).

2.2 AI lifecycle

The Code applies to every part of the design, implementation, procurement, distribution, deployment, maintenance, and decommissioning of an AI system likely to impact children.

The Code is presented in distinct stages. It begins with planning and setting intentions and ends with retirement. In practice, the lifecycle of an AI system evolves through iterative design and when new elements (e.g., new data or instructions) are introduced.

As a result, your consideration of the Code may begin at any point during your AI system's lifecycle. You may move through the stages in a different order and return to different stages on several occasions. For example, when evolving an existing model, a review of outcomes may lead you to reviewing intentions or data inputs. Equally, your testing strategy during later stages might identify shortcomings that make it necessary for you to adjust your operational planning or original intention, for example narrowing or limiting the domains in which your AI system will operate.



AI lifecycle Fig2.

2.3 Context

At all stages of the Code you must consider the context in which your AI system will operate. The Code applies to AI systems across *all* domains and sectors, including, but not limited to, education, health, welfare, science and innovation, creative and entertainment industries, justice, and employment.

The Code is additive and does not replace domain-specific requirements as laid out in law. Context considerations include population and geography (e.g., assessing the number and location of children who will be impacted), security concerns (e.g., opportunities for bad actors to access children or technological maturity), and the current evidence base (e.g., the extent to which current and future capabilities and risks to children are understood). The Code requires a sociotechnical approach, and decisions must be made by those who have the skills and authority to make them.

2.4 Testing and metrics

Assessing AI systems requires a multifaceted approach where each system is subjected to a range of different tests. Testers must have sufficient expertise to design and run tests and to analyse results. Stakeholder engagement (Section 2.5) also provides an important avenue to check and test your proposed approach.

Common testing strategies for AI systems

A/B testing: A form of hypothesis testing where two variants of a piece of software are compared in the field from an end user's point of view.⁴⁷

Red teaming: 'using manual or automated methods to adversarially probe a model or system for harmful outputs and then updating the model and system to mitigate such outputs.'⁴⁸

White box testing: Also called 'structural testing' or 'glass box testing', designing test cases based on the information derived from source code (i.e., the instructions it has been given). The test mainly focuses on the control flow or data flow of a programme, and on verifying that the software is built correctly (verification).⁴⁹

Black box testing: Almost entirely focused on whether or not the system is functional as opposed to knowing the full scope of information about how the system was built. The software tester must not (or does not) have access to the internal source code itself.⁴⁰

Simulation: AI can be used to create systems that simulate the behaviour of virtual agents in a virtual environment.⁵⁰ Simulations can also refer to the process of analysing real-world products or systems through virtual models.

Child-centred testing: Testing strategies that are designed to reflect the unique needs, capacities, and vulnerabilities of children, and that are conducted in a way that upholds their rights under the UNCRC. For example, creating child avatars to replicate typical behaviours would be a safer way to test than allowing children to be exposed to harm.⁵¹

2.5 Stakeholder engagement

Bringing outside voices into all stages of your conformity with the Code increases the likelihood that you will understand the potential risks of the AI system from all perspectives. Stakeholder consultation with children and child development experts prevents adult designers and computer scientists relying on adult assumptions about children.

Your engagement strategy must set out clear objectives – what questions you have, why you need to ask them, and who you need to ask. Once your objectives are clear, consider whether these objectives will need to differ or be adapted for specific stakeholder groups.

Engagement with children needs particular expertise and skills⁵² that may require investing in expertise/skills or engaging with specialist children's rights/children's participation organisations. You may need other expert voices (e.g., domain specialists such as health professionals or teachers). For each group, you must consider how you will ensure full and proportionate representation, including making certain that minoritised or vulnerable groups are represented and that their viewpoints are meaningfully taken into account. Best practice dictates that you provide feedback to participants to let them know what steps you have taken to respond to their input.

2.6 Children's rights and capacities

2.6.1 Children's rights

The United Nations' Convention on the Rights of the Child (UNCRC)⁵³ is the internationally authoritative statement on children's rights and the most widely ratified international human rights treaty in history. It comprises 54 articles that cover all aspects of a child's life. The UNCRC applies to every child across the globe.⁵⁴ It also explains how adults (including business entities) and governments must work together to make sure all children can enjoy all their rights.⁵⁵

CHILDREN'S RIGHTS

Here is a summary of the rights that are most relevant to children and AI systems:

Article 2: The Convention applies to every child without discrimination.

Article 3: The best interests of the child must be a primary consideration in all decisions and actions that affect children.

Article 4: Governments must do all they can to make sure every child can enjoy their rights by creating systems and passing laws that promote and protect children's rights.

Article 12: Every child has the right to express their views, feelings, and wishes in all matters affecting them, and to have their views considered and taken seriously.

Article 14: Every child has the right to think and believe what they choose.

Article 15: Every child has the right to meet with other children and to join groups and organisations.

Article 16: Every child has the right to privacy.

Article 17: Every child has the right to access reliable information from a variety of sources.

Article 19: Every child has the right to be protected from violence, abuse, and neglect.

Article 31: Every child has the right to relax and play.

Article 32: Every child has the right to be protected from economic exploitation and from performing any work that is likely to be hazardous or to interfere with the child's education, or to be harmful to the child's health or physical, mental, spiritual, moral, or social development.

Article 36: Governments must protect children from all other forms of exploitation, for example the exploitation of children for political activities, by the media, or for medical research.

General comment No. 25 on children's rights in relation to the digital environment was adopted by the UN Committee on the Rights of the Child in 2021.⁵⁶ It makes explicit that the UNCRC applies equally in the digital environment and provides additional clarity on its application to the design and deployment of digital products and services.

The usefulness of considering children's rights is illustrated by the following:

- An AI system that wrongly assesses children's academic ability and undermines their right to an education (Article 28) and to non-discrimination (Article 2).
- An AI system that makes it easy for paedophiles to find and abuse children, and threatens their right to life, survival, and development (Article 6).
- An AI system that is not secure, that threatens their right to protection and preservation of their identity (Article 8).
- An AI system that generates inaccurate and misleading synthetic content, that threatens their right to access information (Article 13).

2.6.2 Child and adolescent development

Every child is unique, but more than seven decades of academic research on child and adolescent development has set out the typical capacities, needs, and vulnerabilities of children at different ages and stages of development.

An AI system that makes predictions about a child's behaviour when they are 16 based on behavioural data collected from them when they were eight will not produce fair or reliable outcomes. In the same vein, it would be unreasonable to expect an eight-year-old to independently activate a complaints procedure, while a 16-year-old may choose (or be able to choose) to make a complaint unilaterally without adult oversight. The table in Section 6.1 provides a useful overview of the capacities and vulnerabilities of children at different ages.

All children require consideration, but children in certain age groups may be at further heightened risk in different ways. A frequent mistake is to imagine that younger children are always at greater risk. As children develop, they engage with a greater number and range of digital products and services and are less likely to have adult supervision, so it may be that a 15-year-old is at far greater risk in many more contexts than a five-year-old.

2.7 Diversity and inclusion

Diversity and inclusion strategies must be additive and not used to prevent or block actions that might help most children. For example, if a risk mitigation strategy is shown to protect the majority of children from harm but is less effective in protecting children with learning disabilities, you must solve both issues rather than using its limitations in protecting children with disabilities as an excuse for not putting the mitigation strategy in place for the majority of children.

Intersectional vulnerabilities occur when two or multiple grounds for vulnerability (e.g., age and race or gender) operate simultaneously and interact in an inseparable manner, producing distinct and specific forms of risk of harm.⁵⁷

Your AI systems may be able to support children with different needs, for example by facilitating translation and adaptation for diverse cognitive and linguistic needs, such as tools for sign language recognition, simplified language adaptation, speech-to-text, text-to-speech capabilities, and other accessibility features. Such positive uses of AI are not the subject of the Code but are hugely welcomed by its authors.

2.8 Proportionality

Care has been taken to make the Code straightforward so that it can be followed by small- and medium-sized enterprises (SMEs) as well as those with greater resources.

Assessments of the level of risk posed to children by an AI system must be based solely on the risks created by the AI system and not the size of the organisation. If you cannot build and operate an AI system that conforms with the Code, then you must either narrow its application so it does not impact on children or you must not build and operate the AI system.

Proportionality is not related to the needs of the business, but rather to the level of risk to a child. In practice, this means that the mitigations required by a small tech start-up to ensure its image generator app cannot be used to generate child sexual abuse images must be as comprehensive, fully tested, and robust as for a large company, but equally a large company whose chatbot is focused on supporting parcel deliveries may determine that it poses little or no risk to children other than its (already in place) legal requirements relating to prohibited goods and therefore require no mitigations.

2.9 Role of parents or carers

Many companies outsource safety concerns to parents or carers. Parents or carers have a central role to play, but *they cannot and do not* replace rigorous adherence to the Code, or the basics of conscious design. The Code is for companies to audit their own behaviour and design practice, and adherence to it will provide for products that have actively sought to keep children's needs in mind and implemented their findings by design and default.

PART 3 Criteria

Here are the criteria that you must consider to inform your decisions.

3.1 Developmentally appropriate

The AI system is designed and operated to account for children's differing needs and vulnerabilities at different ages and stages of development by design and default.

3.2 Lawful

The AI system complies with applicable local, national, regional, and international law, rules, and regulations across all domains, including, but not limited to, children's rights, data protection and privacy, child exploitation and abuse, illegal and harmful online content, anti-discrimination laws, consumer protection, intellectual property, health and safety, and education.

3.3 Safe

The AI system does not create or amplify risks to the wellbeing or the physical, mental, and emotional safety of children, including privacy⁵⁸ and security risks.

3.4 Fair

The AI system treats children and their data fairly and creates outcomes that are just and equitable for children.

3.5 Reliable⁵⁹

The AI system functions as expected. Performance and outcomes remain robust over time, including in unexpected or harsh conditions, or when atypical data is introduced. AI systems continue to operate when disruptions occur and recover quickly from incidents. Humans can intervene to take control if required.

3.6 Provide redress

It is easy for children and those who represent their interests to report concerns and to seek actionable and effective recourse and remedy. Complaints relating to children are prioritised and children are kept updated on progress at all stages of the process. The AI system's operators explain their decisions in a way that is easy for children to understand. It is easy to appeal both in-app and without logging in. Emerging concerns or extreme incidents are swiftly and effectively addressed, including by providing an easily reachable human contact.

3.7 Transparent

Stakeholders (including children) have access to adequate and accessible information to have a reasonable understanding of what the AI system does, its impacts, the measures taken to account for the capacities and needs of children, and the efficacy of these measures.

The chain of responsibility for the system, and how design and deployment decisions have been made, are explained.

3.8 Accountable

A continuous chain of human and organisational responsibility is established across the whole AI system's value chain and lifecycle. The results of Code conformity are traceable⁶⁰ and auditable, from start to finish.

3.9 Uphold rights

The AI system upholds children's rights under the UNCRC and General comment No. 25 including their right to life, to participate, and to protection. Inherent in children's right to life is their right to be fully realised as individuals, including meeting their need for agency, connectedness, and purpose. The AI system prioritises children's best interests and takes account of their voices and opinions.⁶¹

PART 4

Common risks to children from AI systems

The purpose of the Code is to identify and mitigate risk of harm to children.⁶² AI systems carry risks for all humans including children – for example loss of human control and malicious attacks or job displacement – but while the Code is cognisant of these often labelled 'existential risks', it is focused on the more quotidian risks that impact on children now and in the near future. These will change as technology changes.

4.1 Unfairness

Unfair outcomes occur when an AI system makes a decision that is wrong because it is discriminatory or inaccurate. The consequences of unfair decisions can be profound, for example children missing out on essential social care, educational/economic opportunities, or fair treatment in the justice system.

Discrimination occurs when bias leads to unfair outcomes for people with similar characteristics. For example, a data set or AI model that includes bias about the qualifications, intellectual abilities, economic status, physical capabilities, or creative interests of girls risks inequitable distribution of education opportunities.

Inaccurate decisions may be based on poor data, incorrect assumptions of the AI model, or specifically in the case of children, because they are assumed to be adults.

4.2 Harmful content and activity⁶³

AI-generated synthetic material can be used to abuse, humiliate, and bully children, to facilitate the spread of misinformation, and to facilitate crime. For example, law enforcement agencies have reported a significant increase in the rise of AI-generated child sexual abuse material.⁶⁴

AI-powered recommender systems play a critical role in the dissemination of harmful content online. They can deliberately increase the visibility of content (human-created or synthetic) (e.g., when a video goes viral), and determine the volume and frequency (dosage) of what a child does and doesn't see (e.g., misogyny vs. an authoritative source, or offering diet advice rather than qualified nutritional advice). Recommender systems also determine other aspects of a child's online experience including how visible they are to others, for example to an unknown adult, who is able to communicate with them, for example

offering friend requests, and the deliberate interventions, for example notifications, alerts, and rewards, that keep them online or persuade them back once they have put their phone down.

4.3 Privacy

Privacy refers to the freedom from unreasonable constraints on the construction of one's own identity,⁶⁵ but can be more simply characterised as the ability to own one's own identity and behaviour including personal and inferred data. AI challenges children's privacy by pervasive data collection, scraping, and processing, which can then be used to profile them (including predicting or provoking their thoughts, feelings, and actions) and to influence or determine their behaviours. AI systems that encourage information sharing and are focused on building networks also undermine children's right to privacy.

Children have less developmental capacity for critical thinking and self-regulation. They also (particularly in their teen years) have a greater desire for peer approval and an increased propensity for risk-taking. Prioritising immediate need over long-term consequences is a feature of childhood that makes children more vulnerable to products and services designed to undermine privacy in order to gather data.

4.4 Security

The confidentiality, integrity, and availability of its training data and/or output data and its onward use are all points of insecurity for AI systems. Security risks can arise from data poisoning, model exfiltration, compromised training data, software vulnerabilities, jailbreaking, or misuse or malicious use of the systems, especially in high-impact domains, for example healthcare, law enforcement, or information integrity.

As for privacy risks, children have less experience or knowledge to recognise security risks, for example phishing or manipulations through deep fakes. Consequently, security risks can lead to more severe immediate, mid- and long-term impacts on children.

4.5 Capture

Digital products and services deploy a range of behavioural design strategies to capture and hold a user's attention. This undermines children's right to agency and exploits their evolving capacities (e.g., they are developmentally less able to self-regulate than adults). Children⁶⁶ form relationships with machines, for example chatbots,⁶⁷ which may impact on their ability to connect with peers and family. This makes them vulnerable to the influence of AI-generated advice, which can be dangerous.⁶⁸ Moreover children learn the skills and competencies required to form and sustain relationships during childhood. Interrupting this development may have long-term effects on relationships in adults.⁶⁹

PART 5

THE CODE

Conforming with the Code is a continuous, iterative process throughout the lifecycle of an AI system. It follows the form of an assessment process so that non-conformity is identified, evaluated, and mitigated in all stages. Progress must be recorded in writing. The criteria set out in Part Three will help guide your decisions. You must also refer to the key considerations in Part Two and the common risks in Part Four.

Preparation: Establishing the oversight and accountability, recruiting, resourcing, and decision-making processes for all stages.

Intentions: Defining the problem that building or deploying an AI system will solve.

Data: Ensuring the quality and appropriateness of the data used to build, train, and operate the AI system.

Development: Designing and training the AI system, including how data sets are prioritised, combined, or weighted to achieve the desired outcome, and what algorithmic techniques are chosen and why.

Deployment: Evaluating the AI system's performance and making the decision to deploy it.

Monitoring: Ensuring the AI system continues to operate as intended following deployment, and addressing issues that arise.

Transparency: Your strategy for providing stakeholders with clear and sufficiently detailed information about the AI system's impact on children, including risks you have identified and the mitigations you have put in place.

Redress: The reporting mechanisms enable children impacted by your AI system to raise issues and concerns.

Decommissioning: The steps you will take to decommission an AI system responsibly.

SNAPSHOT CASE STUDIES

A series of snapshot case studies can be found in Section 6.5. These are illustrative of the sorts of decisions that will need to be made to conform to the Code. In reality, the process of following the Code will require a far more comprehensive risk analysis.

5.1 Preparation

5.1.1 Purpose

The purpose of this stage is to ensure your operational and governance systems are sound.

5.1.1 Outcomes

When you have successfully completed the Preparation stage, you will have:

- (a) Established a process for making decisions, including when and by whom.
- (b) Created a project plan that conforms with the requirements of the Code.
- (c) Provided a realistic estimate of resourcing needs (money, time, and people) that has been approved.
- (d) Assembled a project team with the necessary skills, experience, and competencies.
- (e) Assigned roles and responsibilities to team members and the Executive Leadership for all tasks.
- (f) Made a written record of all the Preparation stage that has been reviewed and signed off (in writing) by the Executive Leadership.

5.1.2 Guidance

Decision making

Significant decisions must be made consciously and transparently by the Executive Leadership including the Chief Executive (or equivalent role). The Senior Accountable Leader (Section 6.1) must have an express mandate to make day-to-day project management decisions and be clear when it is necessary to escalate a decision to the Executive Leadership.

The iterative nature of the design process for AI systems means that decision making must continue throughout the lifecycle of an AI system.

Conformity with the Code requires you to have clear policies and procedures on events or timeframes that trigger a review or report. Throughout the entire AI lifecycle, reporting must be transparent, recorded, and consistent.

Conformity with the Code requires active engagement. Deliberations on how you approach, for example surfacing risk, developing testing strategies, and developing mitigation measures, must be recorded at each stage.

Decision making will require you to exercise judgement and will depend to some extent on context. For example, is it proportional to consider the 'typical' development of all 16-year-olds a bar given that a proportion will be materially less mature than the accepted norm? If you are a company that impacts tens of millions of 16-year-olds, your answer may be different than if your product impacts just a handful of 16-year-olds. In both cases the answer will be to act, but the action may be different. Ultimate decision-making responsibility lies with your Executive Leadership. If you are in doubt, it is better to escalate a decision. You may also wish to seek advice from external experts.

By having a team that covers all skills and disciplines, formalising broad stakeholder engagement, and requiring Executive Leadership to take and be accountable for decisions, many of the most obvious problems will be surfaced and a hierarchy of responsibility established. As the Code becomes normative and in places compulsory, design norms will be established. In the meantime, it acts as a method to create conscious design.

Creating a project plan

The Senior Accountable Leader must create a project plan that is approved by the Executive Leadership. Project plans may incorporate existing risk assessment or risk management systems and processes that support the project team's ability to evaluate conformity with the Code.⁷⁰ All stages of your conformity assessment process must be recorded.⁷¹ Your project plan must include contingency provisions that set out how you will manage unexpected outcomes or issues.

Resourcing your project appropriately

Allocate appropriate resources (both time and money). This includes end-to-end governance and management practices⁷² and human oversight of your AI system at all stages of its lifecycle⁷³ (sometimes termed 'human-in-the-loop'). AI systems typically evolve over time. Conformity with the Code is an ongoing project, and resourcing must reflect this.

Your product launch schedule must account for the time needed to conform with the Code. Team members involved must have sufficient capacity, which may mean reassigning other work and updating individual performance targets.⁷⁴

Lack of resources is not a reasonable excuse for building AI systems that do not conform with the Code. Decisions on allocation of resources require Executive Leadership sign-off.

Assembling your team⁷⁵

Some team members will be core members and some will contribute specific expertise or skills. In addition to their own expertise, team members may need to develop some understanding of subjects that are relevant to the project (e.g., child development and privacy) so that they can collaborate effectively with others.

SMEs and specialist companies may choose to build hybrid teams by bringing in external experts. In all teams (hybrid or in house) a single person may cover more than one role or skill set. Team members may be involved throughout or be called on at certain points, but it is imperative that all of the capabilities are covered, *and every project has a Senior Accountable Leader.*

Here is an example of how a team can be built across internal and external experts. The skills may be divided in many different ways, but in all cases each skill set must be present.

INTERNAL	EXTERNAL
<ul style="list-style-type: none"> • Senior Accountable Leader and Project Manager. • AI systems and risk expertise. • Data, privacy, and security expertise. • Testing and evaluation expertise. 	<ul style="list-style-type: none"> • Child development expertise. • Child rights, ethics, and safety expertise to represent the needs and rights of children as a collective. • Other impacted community rights expertise (including children of these communities). • Stakeholder consultation expertise including consulting with children. • Domain expertise (e.g., health). • Transparency expertise.

For a full list of team members and their responsibilities see Section 6.2.

Qualities and capabilities that high-functioning teams typically have in common include:

- Technical and domain knowledge.
- Experience applying their knowledge in different contexts (knowing what is likely to work).
- Willingness to embrace innovative approaches to problem solving.
- Drive and motivation to achieve goals and strive for optimal rather than convenient outcomes, and resilience when they encounter obstacles or challenges.
- Ability to work efficiently and to minimise waste of project resources.
- Universal behaviours, for example teamwork, leadership, and mutual respect.
- Desire to understand the needs of stakeholders (especially children) as part of a mission to deliver high-quality product and service outcomes.
- Commitment to prioritising children's best interests, even if these are in tension with wider business goals.⁷⁶

Diversity within your team

It is important that your project team reflects a diversity of perspectives and lived experiences. You must address any gaps through training, stakeholder engagement, and consultation. While team members from diverse backgrounds enrich teams, be aware of placing an unfair or unreasonable burden on individuals to speak on behalf of whole communities.

Assigning roles and responsibilities

For teams to operate effectively, everyone needs to be clear on their roles and responsibilities and be accountable for completing them. In a cross-disciplinary team, such as the one required to achieve conformity with this Code, a high level of consultation and input from team members is likely to be needed

throughout. Using a RACI (responsible, accountable, consulted, and informed) matrix for each task is a simple way of ensuring team members know who is responsible or accountable for a task, who must be consulted, and who needs to be informed.

5.2 Intentions⁷⁷

5.2.1 Purpose

The purpose of this stage is to ensure you are clear on what you want your AI system to do and why.

5.2.2 Outcomes

When you have successfully completed the Intentions stage, you will have:

- (a) Carried out an initial exploration of what you want your AI system to do and why (problem statement).
- (b) Assessed your intentions against the criteria (Part Three) to identify and evaluate risk of non-conformity.
- (c) Revised any aspects of your intentions that do not conform with the Code.
- (d) Tested your revised intentions to ensure they now conform with the Code.
- (e) Made a written record of your assessment process and the changes you have made in response that has been reviewed and signed off (in writing) by the Executive Leadership.
- (f) Ensured your project plan aligns with your intentions.

5.2.3 Guidance

Articulating your intentions

Your intentions must be sufficiently precise to enable you to state them. Clearly articulated intentions help identify potential risks. For example, the intention to design an AI system to create filters for user-generated content is not sufficiently detailed if the actual intention is to create filters that 'enhance' attractiveness or enable users to explore what they would look like if they had various plastic surgery procedures.

Intentions may change. Whenever there is a new direction of travel or a significantly different intention, you must assess your intentions again.

Assessing your intentions against the criteria

Once the intention is clear you can evaluate it against each of the criteria. For example, you might:

- Carry out research on similar products or across the domain in which you are working.
- Hold a project team meeting to make sure all team members are aligned.

- Seek the views of children, parents, and relevant experts through, for example, interviews, focus groups, or participatory or co-design processes.
- Take legal advice on the compatibility of the intentions with relevant laws.
- Map potential user journeys for children at different ages.
- Assess the varying ways in which your product may directly or indirectly impact children.

Consider each of the criteria separately and assign a risk level for non-conformity. Give separate consideration to children with heightened vulnerabilities. For example, if your intention is to design an AI system that allows users to take advice from a chatbot, your analysis will differ for younger and older children, or children with learning difficulties.

Revising your intentions

Throughout the intentions stage it is important to ask if the intention could be better fulfilled by a different, less resource-intensive, more cost-effective, or simpler technology. For example, if your AI system automates a simple task and the use cases are extremely limited, you may determine that the risks to children are disproportionate to the benefits, or that GOF AI (Good Old Fashioned AI) is less risky than GenAI (Generative AI).

Even if you determine that the AI system is proportionate, you may still conclude that the risk of non-conformity with one, some, or all the criteria is so high that the project as intended must not go ahead, and the intentions amended or the project abandoned. For example, if you are building an AI system that generates synthetic pornography, you would change your intention to make it fundamental to the design that it can only be accessed by adults. If that was not possible, you would not create the product.

Assessing your revised intentions

If your intentions have changed, you may need to re-assess certain aspects. If they have changed considerably, you may need to re-assess again from scratch including re-testing and/or further consultations.

Recording your assessment process

You must make a detailed record of your process, including feedback, deliberations, testing, methodology, outcomes, and decision making. It also means listing all risks identified in relation to the criteria, including the risks that you have decided fall below the bar of taking any action.

Once this stage is complete the Senior Accountable Leader must submit an Intentions Statement and written account to the Executive Leadership for approval. If the Executive Leadership is not satisfied, they must mandate further steps. The results of these deliberations must be recorded.

5.3 Data

5.3.1 Purpose

The purpose of this stage is to ensure the quality and appropriateness of the data used to build, train, and operate the AI system.

5.3.2 Outcomes

When you have successfully completed the Data stage, you will have:

- (a) Carried out an audit of your proposed or existing data sources/inputs.
- (b) Assessed your data inputs against the criteria to identify and evaluate the risk of non-conformity, including using appropriate testing if necessary.
- (c) Revised any aspect of your data inputs that does not conform with the criteria.
- (d) Tested your revised data inputs to ensure they now conform with the criteria.
- (e) Made a written record of your assessment process and the changes you have made in response that has been reviewed and approved (in writing) by the Executive Leadership.
- (f) Provided in your project plan for ongoing monitoring of your data inputs, including ensuring that data generated by your AI system also conforms with the criteria.

5.3.3 Guidance

AI systems of all varieties contain a series of data points and variables. If the data inputs are problematic, it makes it difficult or even impossible for your AI system to be compliant with the Code. You must assess the quality, integrity, and appropriateness of the data used to build and train the algorithm/model, and be particularly alive to the risk of bias.⁷⁸

If the data needed to train the AI system does not conform with the criteria and creates risks to children that cannot be sufficiently mitigated, the AI system must not be built.

Understanding your data needs

You must be clear on your data requirements, what information you need to train your AI system, and how you will obtain it. For example, if you are building an app that enables football coaches to generate bespoke training programmes, it is likely that you will need performance metrics and indicators for children at different ages, and data about which exercises support training goals. You may need data collected from specific children (case studies) and/or to create synthetic profiles of children at different ages to ensure that your product does not put children at risk, for example by suggesting that young children do training that is too harsh for them.

Sourcing your data inputs

Once you know what data sets you will need, you must consider how they will be sourced. Potential sources include data you have collected, data from publicly available sources, and data sets created and managed by commercial providers. If these sources are insufficient, you may need to supplement them by collecting further data or commissioning a third party to do so on your behalf. You may also decide to create synthetic data sets. In all cases it is your responsibility to ensure that the data you are using is legal and, where necessary, licensed (see Section 2.1).

Irrespective of where you have sourced your data sets, it is your responsibility to ensure that the data collection and management practices conform with the Code – even if you have sourced the data sets

via a third party supplier or open source. You must therefore only use data sets where you can audit the supply chain or trust any assurances you have been given.

Deploying or incorporating foundation and generative models

Evaluating the provenance, quality, and integrity of data when deploying or incorporating generative models poses particular difficulties because of the range of sources and (sometimes) the scale of the data sets on which they are based, the fact that data used to train models is ingested but not stored, the rate at which new (synthetic and human) data is generated, and the challenge of scrutinising data practices of third party suppliers.

Relevant additional questions when incorporating or deploying generative models include:

- Do you share a model spec against which your model has been built and is being deployed?
- If not, what rules or guidelines do you follow to mitigate risks to all users, and specifically to children?
- What steps do you take to account for risks to children impacted by your model?
- Have you included risks to children whose data was used to train the model and children who are not users but who may be impacted by your model?
- Has data used to train the model been obtained with proper legal consent?
- Do you track conformity with agreed standards (whether these are internal or external)?
- If you identify a new or unanticipated risk, who should be notified?
- In the case of an emergency, what is your protocol?

These questions are not an exhaustive list and some are also relevant at the development stage. Your overarching responsibility is to establish whether data is suitable, sufficient, and robust. Your overarching responsibility is to establish whether the data is suitable, sufficient, and robust. Answers to these questions and others that arise from team discussions, deliberation, or external advice or stakeholder engagement must be recorded.

Data input quality requirements

Once you know what data inputs are needed to build and train your AI system, you must evaluate their quality. For example, data used to build, train, test, and validate your AI system must be:

- **Complete:** Your data sets are of sufficient quantity and quality for the use case, domain, function, and purpose of the system. You have sufficient data to generate accurate results for all children, taking into account their age and stage of development and other relevant variables.
- **Balanced:** Your data sets are representative of the diverse groups and characteristics of the children who are likely to be impacted. For example, does your data set include representation of gender, ethnicity, development stage, and socioeconomic status.
- **Unbiased:** Your data sets do not reflect inherent biases against protected or vulnerable groups.

- **Accurate:** Your data sets are reliable, relevant, appropriate, and up to date.
- **Traceable:** The provenance of your data sets is properly recorded, traceable, and auditable.
- **Lawful:** Your data sets must be consistent with data protection laws and other laws.

A NOTE ON DATA HYGIENE

Data hygiene refers to the quality of the data on which the system is built. Data is examined for completeness, bias, and other factors that affect its usefulness for an AI system.⁷⁹ Heightened data hygiene regimes are likely to be needed when using dynamic data, collected and processed in real time for continuous learning.

Make plans at the outset to take data sets, models, and systems out of commission when they are no longer reliable or if they have been shown to create unacceptable risks to children (Section 5.9).

Data hygiene also includes consideration of whether the methods used to collect and process data sets that you are using to build your AI system conform with ethical data-sourcing practices – for example the company has complied with labour laws when employing data processors.

Much of what the Code requires is little more than good data hygiene, which should be an industry norm.

Testing your data inputs

You must decide what testing methodologies are most appropriate to surface and evaluate risk of non-conformity with the Code. At the data inputs stage, testing strategies that are often used include (but are not limited to):

- **Sampling:** Selecting a random but representative sample of the data to check for accuracy and completeness, or stratified sampling to check that representation remains consistent when data sets are sorted or categorised in different ways.
- **Subject matter expert review:** Consulting with subject matter experts to ensure the underlying assumptions are robust. For example, taking the example of the football training programme app, a subject matter expert may be needed to check that ‘normal ranges’ for different groups of children are accurate.
- **Red team testing:** Team members or external experts governed by ethical codes of practice use various methods to test for vulnerabilities in data sets. For example, they may try to unpick pseudonymisation strategies to check for privacy risks if the identity of children can be inferred by those working on the project.
- **Labelling reviews:** Team members review the data labelling strategy for issues, for example bias, duplication, inaccuracy, excessive, or insufficient granularity. Labelling reviews may also surface gaps in data sets that need to be addressed.

Age of users

If your data set includes information about the age of, for example, data subjects, research participants, or end users, you must take steps to understand the accuracy of the data. For example, is it based on tick-box self-declared age, or have further steps been taken to verify or assure the age? The need to establish exact age will be determined by legal requirements, the level of risk, and the presence of prohibited or age-restricted content or behaviours. If the risk to children is high you will need a greater level of assurance. Where it is lower you may determine that a less accurate but more privacy-preserving age assurance strategy is sufficient.⁸⁰ It is always an option to apply the highest standards for all users as a way of providing for the youngest.

Mitigating identified risk of non-conformity

Once you have identified and evaluated data conformity against the Code criteria, you must take steps to address these. Potential mitigations include data cleaning, augmentation, anonymisation, validation, and minimisation, as well as enhancements to your data governance policies and processes. The efficacy of your proposed mitigation strategy must be assessed through further analysis. Once you are satisfied that you have managed risk of non-conformity with the criteria, you must document your process, findings, and recommendations for your review and sign off by the Executive Leadership.

It is inevitable that some products such as those with limited or highly curated data sets will be inherently easier to assess for data hygiene than others. The higher the level of assurance at the outset, the greater the confidence of conformity. Where there is a lower level of assurance, it will be necessary to increase real-time and post-deployment safety strategies, including, but not limited to, frequent testing, introducing a ‘Yellow Card’ system (to allow users to report inaccuracies),⁸¹ automated and human moderation, and more restrictive terms for onward use.

It is not possible to conform to the Code if you knowingly build an AI system using high-risk data and then fail to take mitigations or rely entirely on post-deployment strategies.

A NOTE ON LABELLING AND ANNOTATING PRACTICES

When labelling or tagging data, consider whether social or cultural biases could influence the way it is categorised, or classified, or where automated labelling or annotation could import or replicate historical patterns of discrimination and social or cultural bias.⁸² If third parties undertake the data labelling or annotation on your behalf, ensure that there is appropriate instruction and oversight to guarantee conformity with the Code.

5.4 Development

5.4.1 Purpose

The purpose of this stage is to ensure that the way you design and train your AI system conforms with your Intentions Statement while meeting the criteria in the Code.

5.4.2 Outcomes

When you have successfully completed the Development stage, you will:

- (a) Be clear on the instructions that will drive your AI system.
- (b) Have assessed the instructions against the criteria to identify and evaluate risk of non-conformity using appropriate testing and consultations methods.
- (c) Revised any aspect of your instructions that do not conform with the criteria.
- (d) Tested your revised instructions to ensure they now conform with the criteria.
- (e) Made a written record of your assessment process and the changes you have made in response.

5.4.3 Guidance

Consider the algorithmic instructions⁸³

You must be clear on what instructions will be given on how inputs are to be combined, when, and in what proportions. Considerations include how data sets are prioritised, combined, or weighted to achieve the desired outcome, and what algorithmic techniques are chosen and why. Questions might include:

- Do the instructions need to be adapted for children, or children in certain groups?
- Do you need to provide additional instructions on the types of outcomes it will provide for children, or children in certain age groups (e.g., content restrictions)?
- What is the AI model optimising for?
- Is your optimisation strategy appropriate when the child is likely to be impacted by the outcome?
- How vulnerable is it to prompts and engagement with other users, both general users and those who deliberately target or prey on children?

A NOTE ON SELF-REFERENTIAL AI SYSTEMS

For self-referential or self-learning AI systems, developers must anticipate how the outputs will shape its future behaviours and outcomes. For example, in a recommender feed, if reading, hovering, or clicking one piece of material creates a loop or journey that pushes a user toward similar or more extreme content, that loop may need to be reweighted, mitigated, or balanced for children or children in certain age groups. This is because children by virtue of their development age are more susceptible to persuasion techniques and have less mechanisms to critically assess content or demands of the technology they are engaging with - including advice provided by automated means.

These questions must be answered from a sociotechnical point of view that scrutinises the values, purposes, and interests that shape design and development choices, taking into account the rights and norms of children and childhood. For example, if you have instructed your AI system to learn and predict the geolocation of a service user, this is likely to go against the social norms that adult strangers should not track and monitor children's whereabouts.

Similarly, the AI system must align with the needs of children at different ages. For example, if the AI system is determining the timing of a pedestrian crossing, have those designing or overseeing the instructions anticipated that a primary school child may need more time than an adult to cross? Or, in medicine, has consideration been given to the variation in body weight of pubescent children?

Assessing all your decisions against the criteria will support you to do this process effectively.

Assessing your model against the criteria

Your assessment strategy must be designed specifically to interrogate the impact of the AI system on children and to surface risk of non-conformity with each criteria. For example, if the risk is that your AI system will generate illegal content, you may use proactive detection software to test outcomes. Meanwhile, identifying bias may require you to simulate the behaviours of varying children using avatars and/or discussions with diverse groups of children.

Your approach must be child-centred and consider typical and atypical user behaviours and journeys. You must consider how a child would be impacted if they engage as instructed as well as if they did something predictable but unwanted – for example if a child asks a chatbot to do their homework. Focus groups with children⁸⁴ and consultation with subject matter experts are helpful in surfacing potential risks of non-conformity, which you will then need to assess through testing.

It may be appropriate for your assessment strategy to be developed and run in collaboration with, or with oversight from, independent, accredited, third party experts.⁸⁵

Mitigating identified risks of non-conformity

Where you identify aspects of the AI system design that do not conform with the criteria, you must make changes as needed across, for example:

- product design;
- data management;
- security systems;
- moderation (policies, products and people);
- governance and accountability;
- staff (expertise and training);
- record keeping.

Mitigations by design are the most effective way of changing a child's experience, and should be the primary form of mitigation.⁸⁶

Examples of mitigations you may decide to implement include:

- Age-restricted access to risky features or functionalities powered by the AI system for all children or children in certain age groups.
- Default to protective settings for all children or children in certain age groups.
- Where possible and effective, you may exclude children's data from the AI system. Alternatively, you may exclude specific data points, for example if you have identified risk of non-conformity if chatbots are trained to speak with child-like voices, exclude children's voices from the training set by default.
- Adding a layer of real-time oversight or feedback ensures that a human or AI-driven intervention can be dealt with as it occurs.
- Giving children, or children in certain age groups, additional information to highlight risks created by the AI system at critical points in their user journey. Safety information is supplementary to, not a replacement for, designing your service to be safe or fair.
- Restricting certain prompts or language/image combinations that are known to create harmful material, or use those words to trigger closer oversight.
- Preventing the AI system from engaging with a user who is, or may be, a child.

Once mitigations have been put in place, further testing is required. If risks remain, you may need to cancel or delay launch of the AI system until identified risks are resolved to a manageable level.

5.5 Deployment

5.5.1 Purpose

The purpose of this stage is to decide if your AI system is ready to be deployed.

5.5.2 Outcomes

When you have successfully completed the Deployment stage, you will:

- (a)** Have completed all conformity assessments and testing.
- (b)** Prepared a launch report for the Executive Leadership.
- (c)** Conducted a launch review.
- (d)** Received Executive Leadership approval or reverted to an earlier stage to address issues.
- (e)** Made a written record of the launch review process.
- (f)** Launched your AI system (if agreed).

5.5.3 Guidance

Preparing a launch report(s)

Once you are content that your system is safe to launch, the Senior Accountable Leader must prepare a launch report for the Executive Leadership. This must provide a full account of the AI system development process, including:

- Relevant expertise and qualifications of the team members and their roles and responsibilities.
- A description of the process followed by the team.
- A hygiene report on the data used to build the AI system, including data provenance information about the original data set, and the extent to which machine learning (ML) will inform and impact the data set going forward (data integrity and lifespan).
- Critical assessment, testing, and governance methodologies used and their results.
- Details of external advisors, domain experts, or impacted communities and stakeholder groups consulted, advice received, and how it was taken into account.
- Opinion on non-conformity, including details of any risks that remain fully or partially unaddressed.
- Plans for ongoing monitoring and next stages of iterative development.
- Transparency and reporting strategy.
- Redress mechanisms (Section 5.8), including details of how the complaints will be handled, feedback loops, appeals, human vs. automated decision-making, and how complaints will inform continued risk assessment (e.g., if complaints indicate operational flaws).
- Advice on the anticipated lifespan of the AI system, strategies to manage data decay, scheduled health checks (in addition to continued monitoring), and criteria, which, if met, will trigger the decommissioning and retirement of the system.
- Details of emergency protocols to manage high-risk or catastrophic AI system failures.
- An organisational accountability chart.
- Decisions required from the Executive Leadership before any launch.
- Any outstanding recommendations from the team, including a record of any disagreements or lack of consensus about any aspect of the report.

Launch decisions

The approved report should contain a complete record of your processes and be clearly approved by the Executive Leadership. Commercially sensitive material can be excluded, but not to the extent that it can obscure failure to conform. It is best practice to publish such reviews.

If an AI system is being tested and launched iteratively, it is important to ensure decisions to soft launch or widespread testing are made with oversight from the Executive Leadership across all markets and territories, and that a further review of deployment is introduced between beta rollout and full launch.

5.6 Monitoring

5.6.1 Purpose

The purpose of this stage is to ensure that you have systems and processes in place to monitor your AI system once it has been launched.

5.6.2 Outcomes

To meet your ongoing monitoring obligations, you will:

- (a) Have a plan and capacity for the continued monitoring of your AI system that has been approved by the Executive Leadership.
- (b) Have systems and processes to respond to issues identified through monitoring.
- (c) Run operational and team tests at regular intervals to ensure systems and processes continue to work effectively, and that personnel understand their roles and responsibilities.
- (d) Log monitoring outcomes, including incidents.

5.6.3 Guidance

The requirement to monitor your AI system to ensure conformity with the Code is ongoing. You must take a broad approach to what you monitor, including, for example, unanticipated adverse impacts and harmful outcomes, evolving risks that emerge through continued learning by your AI system (if it is dynamic), and signs of data decay or distributional shift (especially where the speed of deterioration or shift is faster than anticipated).

In high-risk, high-impact AI systems or those involving sensitive or less assured data, monitoring must be constant or frequent and involve independent assessment, and any significant issues identified must be raised for the Executive Leadership. You must have plans in place for a rapid response to unforeseen risks, including requirements for human-in-the-loop systems, kill switches, or other override mechanisms, especially for AI systems with autonomous capabilities.

Incident response protocols

Your launch preparations must include defining a process to follow in the event that your AI system behaves in a way that is unanticipated. The process must include details of how events will be flagged, triaged, managed, and recorded, and how human override will be triggered and achieved if necessary.

Incident reporting

Some industries or domains have requirements to disclose breaches or errors – for example companies must report publicly on data breaches. This is useful for AI systems because those impacted by AI-driven decisions and outcomes have little recourse if they want to challenge and/or trace the decision-making

process. Proactive reporting on, for example, system issues that have led to unfair or unsafe outcomes enable those impacted to manage negative impacts and, potentially, seek redress.

Incident reporting must be provided voluntarily to the regulator(s) (where there is one), and serious or potentially serious incidents must be reported to the relevant authorities, for example government departments, commercial partners, or those communities and children who may be impacted. Where ‘Yellow Card’ schemes or similar formal incident reporting protocols exist, these must be used.

Incident reporting must also work effectively across the supply chain, to all stakeholders deploying the system, and include specific reporting and identification of impacts of the incident for children.

Logging activities

All aspects of your AI monitoring systems must be logged and summarised in plain language with sufficient detail so that any team member or company leader can understand them. Logging systems (including automated logging systems) must monitor and log child-specific data separately.

5.7 Transparency

5.7.1 Purpose

The purpose of this stage is to ensure you are transparent about your AI system and its impact on children or children in certain age groups.

5.7.2 Outcomes

At the end of this stage, you will have:

- (a) Developed a comprehensive transparency strategy that has been approved by the Executive Leadership.
- (b) Developed all aspects of your transparency strategy collaboratively with relevant stakeholders, including children.
- (c) Taken account of the needs and capacities of children at different stages of development and those with additional vulnerabilities.
- (d) Identified ways in which you can provide users with key information about your AI system upfront and throughout the user journey (if your AI system is public facing).
- (e) Continually reviewed and updated your transparency strategy to ensure it is as user-friendly and as useful as possible.

5.7.3 Guidance

Transparency is a key requirement of the Code. It must be possible for external stakeholders to assess the impact and suitability of your system in relation to children – for example to come to a judgement about whether to allow children to use it, to assist academics, or those with oversight or compliance responsibilities (e.g., regulators, auditors, and those operating certification schemes).

Relevant information about your AI system will depend on its purpose and design, but is likely to include:

- Why you decided to build, adapt, or deploy the AI system.
- Why you consider it necessary and proportionate to do so.
- How you considered and responded to the context in which your AI system will operate.
- How you operationalised your plans – in particular, how decisions were taken and by whom.
- What inputs you used.
- Where the data used to train and test the model was found, and how it complies with data protection laws.
- What instructions you included in your model.
- What outcomes your AI system produces.
- Your testing and monitoring protocol.
- Risks identified, how they have been mitigated, and information on remaining risks to children.
- Your external consultation and oversight strategy.
- Your transparency and accountability strategy.
- How those impacted by your AI system can seek redress.
- The history of risk and incidents in relation to children.
- Benefits to children from your AI system.

The level of effective transparency will be determined by, for example, how granular the information is (e.g., whether it is broken down by age of user, country, or region); the time lag on releasing data, which may undermine its utility; the ease of use and the interface; and whether the transparency process has been subject to external, independent oversight or audit.

While the amount and type of information you share is likely to vary depending on the audience (e.g., auditors and regulators may have heightened access rights), it is unlikely to be appropriate to give more privileged access to commercial stakeholders (e.g., advertisers) than to safety stakeholders (e.g., users, non-governmental organisations [NGOs] and academics or regulators).

You must explain your decision-making process when selecting data points for disclosure, including reasons for leaving out key data points, how decisions were made, and by whom. You must also include any caveats or information about known unknowns or anomalies. It is not transparent to provide vast swathes of information that require significant expertise and resource to analyse, and nor is it transparent to provide summaries that obfuscate, triangulate, or omit pertinent information.

Transparency mechanisms

Reports: Produced on a regular basis (e.g., quarterly, bi-annually, or annually) and containing standardised information that can be tracked over time.

Centres: These are (usually) physical spaces where stakeholders (regulators, policy makers, civil servants, journalists, civil society) are given the opportunity to learn more about a company's systems and processes. A centre is only useful if experts are available and stakeholders are able to ask questions rather than listening to a set script.

Information sites: Companies may use websites or in-app information centres to provide information about the AI systems and processes. These are typically aimed at members of the public and provide a top-level overview.

AI model cards: These provide brief explanations of what an AI system does, how it works, the outcomes it generates, and potential risks. The level of detail and usefulness varies depending on the company's approach. They could offer a mechanism for a standardised approach to information sharing.⁸⁷

In-product information: These include buttons such as 'Why am I seeing this video?' or the ability to ask AI systems themselves about how they work.

Registers: These are publicly accessible databases that provide information about AI technologies that are in operation or that have been retired. Registration for these databases may be a statutory requirement or voluntary, and registers may serve the purpose of making specific information about AI systems accessible, for example whether they have been involved in incidents, caused harm, or failed in use.

API access: API (application programming interface) access enables external researchers to interact with an AI model programmatically. This allows academics to receive raw outputs. The utility of an API is determined by the data it provides and who has access to it. To ensure best practice, appoint an informed but independent intermediary to review, approve, and have oversight of requests.⁸⁸ As with any other AI system, APIs must be built and operated in accordance with the requirements in this Code, including security and data protection. They must also take account of academics' heightened ethical conformity obligations so that they can utilise them with confidence.

Audits:⁸⁹ AI audits involve a review of the systems and processes used to design, build, launch, and monitor AI systems to ensure they conform with agreed standards. Audits can be carried out internally or externally. An audit may result in certification, be used as part of internal risk management processes, or as part of best practice efforts. Independent third party audits offer more robust assurance than self-monitoring or privately contracted second party oversight.

Certification schemes:⁹⁰ These use accredited third parties to confirm that an organisation's systems or products conform with a recognised standard or scheme. Standards are typically developed by national or international bodies, for example ISO (International Organization for Standardization) and IEEE (Institute of Electrical and Electronics Engineers). Certification audits

are carried out by accredited bodies that are themselves subject to standards and oversight. For example, the United Kingdom Accreditation Service (UKAS) is the sole national accreditation body for the United Kingdom.⁹¹ In the European Union (EU), CEN/CENELAC (European Committee for Standardisation/European Committee for Electrotechnical Standardisation)⁹² provide the same function across member states. There are also certification schemes for data protection operated under the General Data Protection Regulation (GDPR) in the EU and UK.⁹³

While you are unlikely to have to adopt all these strategies, you must adopt proportionate approaches to ensure that users, regulators, commercial partners, independent researchers, and those who are impacted have a fair and transparent understanding of your AI system.

Ensuring transparency information is accessible to children

You must also ensure that the way you communicate information is clear for children. Ethical UX research⁹⁴ techniques should be used to identify how best to position information at different points of a child's user journey.

All children in all contexts require information in a form they can understand, which may require formats aimed at different age groups, in different languages.

Your responsibility to be transparent is ongoing. You must be proactive in your efforts to continually enhance children's understanding. This includes sharing your transparency plans with external stakeholders so that they can tell you what information they would find most helpful and seeking – and responding to – feedback.

5.8 User reports and redress

5.8.1 Purpose

The purpose of this stage is to ensure you have effective report and redress systems and processes in place to facilitate user reports from or on behalf of children.

5.8.2 Outcomes

At the end of this stage, you will have:

- (a) A comprehensive user reporting strategy that takes account of the needs and capacities of children at different stages of development and those with additional vulnerabilities.
- (b) Created a way for parents, carers, and teachers to report on behalf of children that does not require being logged into or registered to your product or service.
- (c) Co-created your user reporting strategy with relevant stakeholders including children.
- (d) Signed off your user reporting strategy with the Executive Leadership.

(e) A protocol in place to inform regularly or in extremis the relevant authorities about emerging risks or incidents.

(f) A plan to periodically review and update your user reporting strategy.

5.8.3 Guidance

Redress mechanisms enable those who use AI systems to complain about or challenge the impacts of the AI systems. While user reporting is now a common aspect of service design, these typically facilitate complaints about inappropriate content or activities rather than the impact of AI systems (e.g., recommender systems).

The ability to make a report is an important aspect of service delivery in both the public and private sector, since it may be the only way that an individual can confirm that they have been impacted by an AI system or raise concerns about negative outcomes.

Creating comprehensive user reporting systems

User reporting systems must:

- Be prominent. Include the option to report AI systems, alongside reporting mechanisms for related service issues (e.g., privacy, security, and harmful content and activities).
- Be promoted. Highlight your reporting systems to users during onboarding and at relevant moments in their user journey. For example, if a user uses a 'Show me less of this kind of video' function when watching an AI-generated video in their feed, you must let them know they can make a report about what they are being offered.
- Be intuitive. It must be quick, easy, and obvious for children to access and activate.
- Be user-centred. If you offer users options for categorising their report, these must reflect their experiences and language rather than aligning with your internal categorisations. There should always be an 'other' option in the report menu.
- Be child-friendly. The language and design must appeal to children.
- Be accessible. Consider the additional needs of children with accessibility challenges.
- Be frictionless. Identify and remove any barriers to children making a report, for example asking children to provide unnecessary information, or asking them to diagnose or categorise their issue when they don't have the language and to do so.
- Be privacy-respecting. Children can raise a report themselves and do not have to obtain consent from a parent or carer to do so.
- Allow for adult support where wanted. Children can include parents or carers in a reports procedure if they choose to.
- Enable adults to raise reports on behalf of an individual child or on behalf of a wider group of children. This means providing report mechanisms in-product and in your help centre. You must not require reporters to have or create an account in order to make a report.

- Be transparent. Automate or regularly report on emerging concerns or thematic complaints to the relevant authorities.

Reflecting the unique character of your service in your reporting systems

Your reporting system must align with the way children use or may be impacted by AI systems on your service. For example, children may wish to complain about use of AI systems to decide what ‘friends’ recommendations they receive on a social media platform; whether they get into a college or university admissions; the price of their car insurance; the predictions made about their academic potential; the content they see; the amount of time they are encouraged to spend on a service; the advertisements they are shown; or the team they are assigned to on a gaming site, among other things.

You must provide an option for them to describe their concern (e.g., ‘I have a different problem’).

Testing your user reporting system with stakeholders including children

Co-creating and testing user reporting pathways with children means you can be confident you understand their needs and the processes you design work for them at different ages and with heightened needs and vulnerabilities.

User reporting systems should also be tested, and validated by, independent experts. Doing this maximises the likelihood that the reporting mechanisms will operate as expected and give you the opportunity to troubleshoot any unanticipated issues.

Once launched, your user reporting systems must be monitored and regularly reviewed to ensure they continue to operate as expected.

Responding to user reports

It is not sufficient to give children the opportunity to report issues. You must also ensure you provide a timely and clear response.

This means you:

- Provide confirmation that you have received their report and indicate when they can expect a response and who to contact if they haven’t heard back or circumstances change.
- Prioritise reviewing reports received from children.
- Tell children what decision you have made, what steps you will take to resolve the issue, and, if required, who and how to escalate any complaint.

Implementing changes in response to complaints

Once an issue has been identified, you must respond. It may be that the report is isolated and specific to the child who has raised the concern. Or a complaint might point to a wider issue. When this happens, you have a responsibility to investigate and solve systemic problems surfaced for all users, and not just for the individual who has reported it.

If an AI system is operating in breach of the criteria, take proactive steps to let other children who have or may have been impacted know what has happened, what it means for them, and what steps you have taken to ensure that it doesn’t happen again.

Consider a range of possible responses to a report that might include, for example, removing a user’s data from a data set, providing information to the user, improving the information you provide publicly about your AI system to all users, reviewing similar decisions and adjusting outcomes, redesigning weighting or prompts, temporarily withdrawing or retiring the AI system, carrying out further testing, establishing closer monitoring procedures, or notifying regulators or other oversight bodies.

Notify your supply chain where relevant.

Opportunities to enhance transparency

Complying with requests for information or questions about your processes and systems as well as requests for action.

Including information about your report and redress processes in your transparency reporting enables others to gain a better understanding of the impact of your AI system on children.

5.9 Retiring and moving on

5.9.1 Purpose

The purpose of this stage is to assist you in planning to decommission your AI system, including predicting and monitoring its likely retirement date.

5.9.2 Outcomes

At the end of this stage, you will have:

- (a) Agreed the criteria against which you will assess life expectancy and the cadence at which it will be reviewed.
- (b) Carried out a preliminary review of your AI system’s life expectancy.
- (c) Conducted a decommissioning impact assessment for a planned and emergency retirement of your AI system.
- (d) Been clear what steps you will need to take to retire your AI system, and the resources (people, time, and money) required to complete the process.
- (e) Have emergency protocols in place in the event that it becomes necessary to retire your AI system at short notice.
- (f) Secured written approval from the Executive Leadership of the retirement protocols, assessment, and planning.

5.9.3 Guidance

When thinking about whether an AI system is still operating effectively, it is important to consider whether it continues to meet standards in the Code and not simply whether it is operationally or commercially sufficient. Once the AI system has deteriorated to the extent that it is no longer compliant with the

criteria in the Code, it should be decommissioned and retired, or subject to a redesign and relaunch by reconsidering each of the stages. This process must be planned from the outset to ensure that consequences of retirement are anticipated and addressed in advance.

Decommissioning is an often little considered stage in an AI system's lifecycle. This is reflected in global standards, protocols, and guidance that tend to focus on building and deployment. The National Institute of Standards and Technology's (NIST) Trustworthy & Responsible AI Resource Center⁹⁵ provides helpful guidance on decommissioning.

When should an AI system be decommissioned?

An AI system may be decommissioned because:

- The accuracy and performance of the AI system has degraded to the point where it no longer conforms with the Code or will soon fall below Code requirements.
- You have launched a new version, which means the old version is redundant.
- The data used to train the AI system has become outdated or unreliable.
- New regulations will come into force that will render your AI system non-compliant.
- You have become aware that your AI system is producing unsafe or unfair outcomes.
- Security protocols are outdated, making it vulnerable to cyberattacks.
- Maintenance costs have increased to a point where operating the AI system is no longer commercially viable.

Decommissioning is only one possible response to the issues set out. You might also decide to repair or repurpose your AI system. If so, you must ensure that the AI system continues to conform with the Code.

Evaluating whether an AI system should be decommissioned

Evaluating whether an AI system has reached the end of its useful lifespan (or will soon do so) should be informed by the criteria set out in Part Three.

Questions to consider include:

- **Data quality:** Are the underlying data sets and the outcomes generated sufficiently accurate and reliable?
- **Algorithmic integrity:** Are the instructions driving your AI system still appropriate and effective? Have algorithmic assumptions become outdated or out of step with public opinion on, for example, bias, discrimination, or age-appropriate content standards?
- **Regulatory standards:** These may change or be introduced, which makes your AI system unsustainable. Although conformity with the Code makes it more likely that your AI system will comply with local and regional laws, you must continually monitor new and incoming legal requirements to understand whether your AI system complies.

Carrying out a decommissioning impact assessment

Before decommissioning your AI system, you must carry out an impact assessment to identify and mitigate potential negative consequences for children. Your assessment may need to be repeated if you have to decommission the AI system quickly in response to an emergency.

As part of your impact assessment, you must consider:

- The impact on children if they cannot access the AI system, or if services or products they rely on are impacted by the decommissioning.
- The impact on children if other systems, services, or products that you control are affected by the decommissioning of the AI system.
- The impact on children if systems, services, or products operated by others in your onward supply chain are affected by the decommissioning of the AI system.

Impacts must be assessed against the Code criteria. For example, if you are decommissioning an AI system that offers personalised learning courses to children, it is likely to be unfair and incompatible with their right to education if you do not give them sufficient time to finish their course or extract evidence of completion unless you provide an alternative, equivalent product or service and make it easy for them to transfer over.

Your impact assessment must evaluate the level of risk and include proposed mitigations. For example, if you identify as a high risk that the security of sensitive personal data about children in the data sets used to continuously train your AI system will deteriorate once the AI system is decommissioned, you must put in place a plan to manage this risk (e.g., arranging for the secure and irreversible deletion of the data set).

All members of your project team will need to contribute to all stages of the decommissioning impact assessment, including developing a testing strategy to surface and evaluate non-conformity, reviewing and interpreting test results, and agreeing mitigations. Ultimately it is the responsibility of the Executive Leadership to manage this process effectively.

Planning your decommissioning process

Once you have completed your decommissioning impact assessment and it has been signed off by the Executive Leadership, you will be able to plan your decommissioning process. Your team's Project Manager is responsible for creating a project plan, timeline, and budget in consultation with team members and the Senior Accountable Leader. The Executive Leadership will agree a preliminary version of this plan as part of the launch review, but they will also need to sign off on the final version when the time comes.

Your launch plan may include:

- How you will make sure AI systems are securely decommissioned to prevent unauthorised access or misuse.
- Procedures for data sanitisation and secure disposal of model components.
- Transition plans so that children have sufficient notice that an AI system that impacts them will no longer be available and are given help with any transition requirements (e.g., updating an app on the app store).

If you anticipate that the decommissioning of your AI system will have negative consequences for some children (e.g., those who cannot upgrade to a new system or who require a key feature of the old system), you may wish to consider keeping a legacy version available, provided this can be done in conforming with the Code.

Supply chain

Consider the onward supply chain and ensure that services that use or rely on your AI system are aware of your plans to retire your AI system in advance. It is your responsibility to ensure full retirement across your supply chain. (This must be anticipated in supply agreements from the outset.)

If you are deploying an AI system and are notified by the developer that the system has been retired or substantially changed, you must respond to the notification and consider the consequences for your own AI system and take action.

Repurposing some or all of your AI system

Repurposing, patching, and adapting systems based on decaying data or flawed models creates and amplifies risk. Once an AI system has been retired or is no longer operating within safe boundaries, it may be appropriate to re-use some of its constituent parts. This should be done only if conformity with the Code is guaranteed. It is likely that any such action will require the full conformity process to use all or part of the AI system in another setting.

PART 6

Further context & definitions

Part Six includes information to help those using the Code to understand it better or to delve into greater detail.

6.1 Childhood development

This table is inevitably a generalisation and does not capture all the nuances and diversity of children's experiences, attitudes, and interests. However, while child development is neither entirely linear nor homogenous, and there is no universal blueprint for 'standard' capacities, it reflects years of academic research and current regulatory guidance.⁹⁶



0-5 YEARS-OLD⁹⁷

Significant numbers of children are online from the earliest of ages.⁹⁸

In the UK in 2023, 27% of three- to four-year-olds had their own mobile phone.⁹⁹ They are predominantly engaged in adult-guided activities, playing within 'walled' environments, or watching video streams. Children play on their parents' devices, which may not be set up with child-specific profiles.

For children up to the age of two, autoplay and algorithmically driven recommendation functions can lead to passive use that takes attention away from the necessary activities required for development, for example movement, free play, and learning emotional cues.

At ages three to five, children start to develop the ability to 'put themselves in others' shoes' (theory of mind) but are easily fooled by appearances and tend to believe what they see. They are developing friendships, although peer pressure is relatively low, and parental or family guidance or influence is key. They are learning to follow clear and simple rules, but are unlikely to have the cognitive ability to understand or follow more nuanced rules or instructions.

Play (role-play, messy play, structured play, free play etc.) continues to have an essential role in their social, emotional, physical, and cognitive development. They have limited capacity for self-control or ability to manage their own time online.

Excessive engagement with digital content may lead to decreased engagement with their physical environment. For example, children are stationary, unaware of their surroundings, not talking or interacting with others.



6-9 YEARS-OLD

Children in this age range are likely to have their own device, although use of parents' devices is still common. They are increasingly using devices (tablets, phones, consoles, and connected toys) independently. They may engage enthusiastically with voice-activated devices (e.g., smart speakers). Children often use digital gaming and creative-based activities and video-streaming services. Some may be experimenting with social media use, either through social aspects of digital games, through their parents' social media accounts, or by setting up their own social media accounts. They may relate to and look up to influencers.

They may be absorbing messages from school about digital safety and be developing a basic understanding of privacy concepts and some of the more obvious digital risks. They are unlikely, however, to have a clear understanding of the many ways in which their personal data may be used, or of any less direct or obvious risks, for example their digital footprint, commercial profiling, or being drawn down the 'rabbit holes' that their digital world may expose them to. Limited critical understanding can mean that neither veracity of information nor its purpose is questioned and properly understood, particularly if they are making friends with either humans or chatbots that are interested in them.

They are becoming more socially sophisticated. The need to fit in and be accepted by their peer group becomes more important. Awareness that their social status can be influenced by their skills and acquisitions increases. Collecting (e.g., cards, figures, skins) becomes a powerful way of demonstrating this status, and can also be a source of comfort.

The need to fit in with their peer group becomes more important towards the end of this age range, so they are increasingly susceptible to peer pressure. However, home and family still tend to be the strongest influencer. They still tend to conform with clear messages or rules from home and school, but often fill any gaps with explanations of their own or come up with protective strategies that aren't as effective as they think they are.

Habits formed before the age of seven are hard to change later in life, so from birth to five and six to nine are crucial ages to develop a healthy relationship with digital activities.

The absence of a common culture on introducing devices means questions of autonomy vs. oversight may be a source of family tensions.



10-12 YEARS-OLD

Children are much more likely to be given their own device and to ask to use age-restricted apps because others in their friendship group have them. Not being given access to apps used by their peers may be a source of anxiety and family tensions. Parents and children may be unaware that apps and games that children use are mixed age spaces in which children are accessible to unknown adults or far older children.

There is a shift towards use of the digital environment to explore and develop self-identity and relationships, to expand and stay in contact with their peer group, and to 'fit in' socially. Use of social networking functions or services increases. Self-esteem may fall as children compare themselves to others and strive to present an acceptable version of themselves online.

Attitudes towards parental rules, authority, and involvement in their digital activity may vary considerably, with some children relatively accepting and others seeking or even demanding higher levels of autonomy. However, parents and family still tend to be the main source of influence for children in this age range. Children are moving towards more adult ways of thinking, but have limited capacity to think beyond immediate consequences, and are particularly susceptible to reward-based systems (likes, shares, comments, gifts), and tend towards impulsive behaviours. Parental or other support still tends to be needed, if not always desired, even if it must be in a less directive way than for younger children.

Children are developing a better understanding of how the digital environment operates, but are still unlikely to be aware of less obvious uses of their personal data or how the design of services is deliberately compulsive. They may express being online as a need, or have feelings of shame or embarrassment when they struggle to self-regulate the amount of time they spend on their devices. The volume of group chat messages, notifications, and content can be overwhelming.

Mixed messages may be an issue as schools require greater engagement with devices for receiving messages and uploading homework, but then warn of the dangers, particularly of bad actors. While schools may discuss the impact of compulsive loops, advice on phone hygiene (in which alerts, buzzes, and unused apps are switched off or discarded) may be more limited and inadequate to address these issues.



13-15 YEARS-OLD

The use of social media functions and apps is widespread, and changes as new products enter the market. Posting, gaming, and video - and music - streaming services remain popular. The use of new services that parents aren't aware of or don't use is popular, as is the use of language that parents may not easily understand.

The need for identification with their own peer group and exploration of identity and relationships increases, and children are likely to seek greater levels of independence and autonomy. Again, they may seek to emulate influencers at this stage or the more powerful social actors in their network.

Children may tend toward idealised or polarised thinking and be susceptible to negative comparisons of themselves with others. Impulsivity and compulsion to seek rewards remains high. There is the potential for emotional contagion (positive and negative) and distress as they become aware of discrepancies between the 'ideal' presented by others online and their own reality. This may include a negative body image.

They may reject or distance themselves from their parents' values, or seek to actively flaunt parental or digital rules, and show more sensitivity to risk. Some children become more risk-averse while for others the desire to seek out risk increases.

The recent surge of ‘chatbot’ friends has added a new dimension. These have been found to engender trust but may reinforce negative feelings and, in some cases, push children to harmful states or activities. Many children cite the lack of other opportunities to spend time with their friends as a driver to be online. They also complain of parents’ hypocrisy and inattention as parents spend time on their phones in the family home while telling them to put theirs down.

Children may still look to their parents to assist if they encounter problems online, but some may be reluctant to do so if they have broken agreed rules or believe their device will be taken away. They may overestimate their own ability to cope with risks and challenges arising from digital behaviour and relationships, and may benefit from signposting towards sources of support, including, but not limited to, parental support.



16-17 YEARS-OLD

Children in this age group often look and behave in adult ways, but they are still developing cognitively and emotionally, and cannot be expected to have the same resilience, experience, or appreciation of the long-term consequences of their digital actions as adults.

They are still influenced by their peers but are likely to have found their social niche and to prioritise more intense relationships.

Their technical knowledge and capabilities may be better developed than their emotional literacy or their ability to handle complex personal relationships. Their capacity to engage in long-term thinking is not yet fully developed, and some still tend towards risk-taking or impulsive behaviours and are susceptible to reward-based systems. They may have developed coping strategies for their feelings of being overwhelmed, addicted, failing to measure up, or doom scrolling (endlessly engaging with empty or poor-quality content), but often feel shame when those strategies don’t work.

Children typically have complete authority over their screen use. Parental support is more likely to be viewed as something that they may or may not wish to use rather than as the preferred or only option. Signposting to other sources of support in addition to parental support is important. Young people often express the view that they have not received adequate explanations and understanding of the way in which their emotions and activities have been orchestrated by the products they engage with. Many feel that the safety measures are focused on adult concerns, and also that adults have not adequately understood or protected them from risk of harm.

Children may note that they would ‘get offline’ if all their friends did, citing the lack of opportunity to meet friends in real life.

They may prepare to enter the adult world with a significant digital footprint that may impact on their ability to access education and employment opportunities. For some, this legacy may include non-consensual intimate images. The impact of the normalisation of children’s access to harmful and inappropriate content, including pornography, may have consequences for some children’s ability to form and maintain healthy relationships in the long term.

6.2 Team member roles and responsibilities

Some roles may be internal and some external consultants may be required. One person (whether an internal team member or external expert) may cover more than one role (especially in smaller organisations). Some team members will only be involved in a single aspect of the process. All capabilities must be covered, and every project must have a Senior Accountable Leader.

SENIOR ACCOUNTABLE LEADER

This person is part of the organisation’s Executive Leadership. They hold responsibility for the project and must have a clear reporting line to the Chief Executive Officer (CEO) (or equivalent).

Their responsibilities include:

- Establishing corporate commitment to conforming with the Code (e.g., by including it in the organisation’s business strategy).
- Ensuring organisation-wide efforts to conform with the Code are appropriately resourced and prioritised.
- Advocating for changes required as a result of the project, including those that are in tension with the business’s commercial strategy.
- Resolving any issues relating to implementation or disagreements within the team as they arise.
- Championing the Code to cross-organisational stakeholders so that they understand its importance, to ensure cooperation and support.
- Overseeing preparation of, and sign off on, the launch report plus any interim or subsequent reports to the Executive Leadership.
- Briefing the Executive Leadership in a transparent manner, including other team members where they have additional skills or information.
- As appropriate, providing updates to company board members and shareholders.
- Ensuring all decisions are made at the appropriate level.
- Acting as the company’s representative when asked to account for conformity with the Code to external stakeholders, including policy makers, regulators, and customers.
- Embedding a culture of continued conformity with the Code throughout the organisation and the lifecycle of the organisation’s AI systems.

PROJECT MANAGER

This person is responsible for ensuring the project runs smoothly, that it is staffed appropriately, and completed within budget and on schedule. They report to, and work closely with, the Senior Accountable Leader throughout the project. In a small company the Project Manager may also be the Senior Accountable Leader, but it is a key function of the Senior Accountable Leader that they have adequate experience,

resources, and authority to conform with the Code without being diverted by other interests, either on their time or their decision making.

Their responsibilities include:

- Working with the Senior Accountable Leader to scope the project and recruit expert team members.
- Working with the Senior Accountable Leader to define team member roles and responsibilities including creating a RACI (responsible, accountable, consulted, and informed) matrix and keeping it up to date.
- Drawing up the project plan, timeline, and budget, and reviewing and updating it throughout.
- Coordinating team meetings and sending regular updates.
- Liaising between different team members and with cross-functional teams whose input is required.
- Securing additional subject matter expertise when the team identifies a gap in their knowledge or where the Senior Accountable Leader concludes it is appropriate to obtain independent advice.

AI SYSTEMS EXPERT

This person has oversight and understanding of all AI systems deployed across the organisation.

Their responsibilities include:

- Ensuring all team members have sufficient understanding of how AI systems work and how they are deployed across the organisation.
- Answering subject matter questions from colleagues and external advisers.
- Listening to other team members to understand concerns and potential solutions and working collaboratively towards solutions (even if they are technically more complex to implement).
- Documenting technical decisions made and the logic for these decisions.
- Liaising with technical experts within the organisation and securing engineering and other human resources needed for all stages of the project.
- Providing input into identifying risks in intentions, inputs, instructions, and impacts.
- Providing input into testing strategies and review of outcomes.
- Providing input into evaluating proposed mitigations.
- Providing input into monitoring and oversight strategies.

AI RISKS EXPERT

This person is an expert in the full range of risks that AI systems pose and current best practice approaches

to risk mitigation. They work closely with the Age-Appropriate Expert to understand and articulate how these risks manifest for children, and how solutions may need to be adapted to reflect the rights, needs, and capacities of children at different ages and stages of development.

Their responsibilities include:

- Ensuring all team members have sufficient understanding of AI risks and current best practice solutions.
- Answering subject matter questions from colleagues and external advisers.
- With the Age-Appropriate Lead, describing how risks manifest for children and the ways in which a child-specific context alters best practice approaches.
- Drafting the conformity review framework.
- Providing input into testing strategies and review of outcomes.
- Providing input into evaluating proposed mitigations.
- Providing input into monitoring and oversight strategies.
- Providing input into reports for the Executive Leadership.

AGE-APPROPRIATE EXPERT

This person is an expert in the rights, needs, and capacities of children at different ages and stages of development. They have experience in applying their subject matter expertise to the design and oversight of digital products and services including AI systems. They understand how risks across areas such as safety, fairness, and transparency manifest for children generally, and also the way risks change at different ages and with additional vulnerabilities. They are skilled at collaborating and communicating with colleagues across the organisation, including engineering, research and development, legal, product team, and management.

Their responsibilities include:

- Ensuring all team members have sufficient understanding of child development, the way in which it impacts risk from AI systems, and current best practice solutions.
- Answering subject matter questions from colleagues and external advisers.
- Providing input into testing strategies and review of outcomes.
- Providing input into evaluating proposed mitigations.
- Providing input into monitoring and oversight strategies.
- Providing input into reports for the Executive Leadership.
- Evaluating technical, operational, and design decisions with regard to age needs.

CHILD RIGHTS AND VOICE EXPERT

This person is responsible for ensuring children's rights under the United Nations Convention on the Rights of the Child (UNCRC) and General comment No. 25 are upheld and their best interests prioritised. They also ensure that children's views are taken into account throughout the process. They are an expert in participatory research (quantitative and qualitative). They are committed to ensuring children are active participants in decisions that impact them, and that outcomes of consultation exercises truly reflect their contribution.

Their responsibilities include:

- Ensuring all team members have sufficient understanding of children's rights and the way they impact design of, and decisions about, AI systems.
- Answering subject matter questions from colleagues and external advisers.
- Running research programmes with children.
- Conducting desktop research to gather research done by others that is relevant (this is likely to be more appropriate for smaller organisations).
- Providing input into testing strategies and review of outcomes.
- Providing input into evaluating proposed mitigations.
- Providing input into monitoring and oversight strategies.
- Providing input into reports for the Executive Leadership.

AI TESTING EXPERT

This person is an expert in designing and running multifaceted testing strategies to surface and evaluate risk, and to test the efficacy of mitigation strategies throughout the lifecycle of an AI system. They work closely with the Age-Appropriate Expert, Children's Rights and Voice Expert, and AI Risk Expert to ensure the testing strategy is fit for purpose and tailored to the specific needs and capacities of children.

Their responsibilities include:

- Ensuring all team members have sufficient understanding of how to test AI systems.
- Answering subject matter questions from colleagues and external advisers.
- Designing and implementing all aspects of the testing strategy for the AI system at all stages of the lifecycle.
- Liaising with external experts as needed.
- Ensuring testing results are validated and reliable.
- Providing input into evaluating proposed mitigations.

- Designing monitoring and oversight strategies.
- Providing input into reports for the Executive Leadership.

DATA SET EXPERT

This person is an expert on data hygiene practices. They may well work within the AI Systems Lead's team.

Their responsibilities include:

- Ensuring all team members have sufficient understanding of data set management.
- Answering subject matter questions from colleagues and external advisers.
- Leading on the review of data sets that have or will be used to build, train, and test the AI system.
- Assessing and advising on all aspects of data hygiene.
- Certifying and auditing supply chain data.
- Providing input into testing strategies and review of outcomes.
- Providing input into evaluating proposed mitigations.
- Providing input into monitoring and oversight strategies.
- Providing input into reports for the Executive Leadership.

PRIVACY EXPERT

This person is an expert in privacy law and data rights as they manifest in AI systems. They understand children's heightened privacy rights. They are an expert in local standards, global frameworks, and best practice approaches. They work closely with the Age-Appropriate Expert, and are able to spot circumstances where they need further information to advise.

Their responsibilities include:

- Ensuring all team members have sufficient understanding of privacy and children's right to heightened protection of their personal data.
- Answering subject matter questions from colleagues and external advisers.
- Ensuring children's right to privacy is considered and upheld at all stages of the AI system's lifecycle.
- Collaborating with the AI Test Expert to ensure the testing strategy is effective in surfacing, evaluating, and monitoring privacy risks to children.
- Providing input into testing strategies and review of outcomes.

- Providing input into evaluating proposed mitigations.
- Providing input into monitoring and oversight strategies.
- Providing input into reports for the Executive Leadership.

SECURITY EXPERT

This person is an expert in security risks as they manifest for AI systems. They understand children's heightened exposure to security risks. They are an expert in local standards, global frameworks, and best practice approaches.

Their responsibilities include:

- Ensuring all team members have sufficient understanding of an AI system's security.
- Answering subject matter questions from colleagues and external advisers.
- Ensuring the organisation's AI systems are secure for children.
- Collaborating with the AI Test Expert to ensure the testing strategy is effective in surfacing, evaluating, and monitoring security risks to children.
- Providing input into testing strategies and review of outcomes.
- Providing input into evaluating proposed mitigations.
- Providing input into monitoring and oversight strategies.
- Providing input into reports for the Executive Leadership.

TRANSPARENCY EXPERT

This person is an expert in best practice strategies to increase transparency of AI systems. Their knowledge includes reporting, user information, and API (application programming interface) access. They work closely with the Age-Appropriate Expert to ensure transparency efforts are child-friendly.

Their responsibilities include:

- Ensuring all team members have sufficient understanding of transparency.
- Answering subject matter questions from colleagues and external advisers.
- Ensuring the organisation's approach to transparency conforms with the Code.
- Managing requests for information, and ensuring that data is collected and presented in a way that facilitates scrutiny and oversight.
- Providing input into testing strategies and review of outcomes.

- Providing input into evaluating proposed mitigations.
- Providing input into monitoring and oversight strategies.
- Providing input into reports for the Executive Leadership.

DESIGN LEAD

This person is an expert in the responsible design of digital products and services. They work closely with the Age-Appropriate Expert to ensure that design decisions adhere to age-appropriate design principles.

Their responsibilities include:

- Ensuring all team members have sufficient understanding of UX Design.
- Answering subject matter questions from colleagues and external advisers.
- Identifying where and how children are likely to engage with the AI system when using a digital product or service.
- Leading on the design of user reporting systems that are child-friendly and intuitive to use.
- Ensuring advice (including warnings) about AI systems is proactively surfaced to children and their parents or carers.
- Ensuring safety, security, and privacy settings are easy to find and understand.
- Auditing deceptive or persuasive design strategies that are incompatible with children's best interests or evolving capacities.
- Providing input into testing strategies and review of outcomes.
- Providing input into evaluating proposed mitigations.
- Providing input into monitoring and oversight strategies.
- Providing input into reports for the Executive Leadership.

6.3 Definitions

This section includes definitions categorised by:

- A. Children and childhood;
- B. AI;
- C. Data;
- D. Safety by design.

Not all terms in this section appear in the Code. Some are included because they are helpful in providing the wider context.

A. Children and childhood

AGE APPROPRIATE

A term used in legislation and regulation that (outside the US) is associated with the United Nations Convention on the Rights of the Child (UNCRC) and its provision that children must be treated according to their evolving capacity. Title 42 of the US Code,¹⁰⁰ which covers public health, social welfare, and civil rights, defines ‘age or developmentally-appropriate’ as ‘activities or items that are generally accepted as suitable for children of the same chronological age or level of maturity or that are determined to be developmentally-appropriate for a child, based on the development of cognitive, emotional, physical, and behavioural capacities that are typical for an age or age group.’¹⁰¹

AGE ASSURANCE¹⁰²

An umbrella term for both age verification and age estimation solutions. The word ‘assurance’ refers to the varying levels of certainty that different solutions offer in establishing an age or age range.

AGE ESTIMATION

A process that establishes that a user is likely to be of a certain age, fall within an age range, or is over or under a certain age. Age estimation methods include automated analysis of behavioural and environmental data; comparing the way a user interacts with a device or with other users of the same age; metrics derived from motion analysis; or testing the user’s capacity or knowledge.

AGE VERIFICATION

A system that relies on hard (physical) identifiers and/or verified sources of identification, which provide a high degree of certainty in determining the age of a user.

BEST INTERESTS OF THE CHILD

The UNCRC sets out a set of immutable rights to which all children are entitled. Some rights are protective¹⁰³ and some participatory.¹⁰⁴ The onus is placed on those interpreting their obligations to decide how to balance them in different circumstances, taking into account local laws. In making this decision, the overriding principle is that the child’s best interests is a primary consideration.¹⁰⁵ General comment No. 25 provides helpful guidance on how a child’s primary UNCRC rights apply in the digital environment, but makes clear that the ‘best interests of the child is a dynamic concept that requires an assessment appropriate to the specific context.’¹⁰⁶

CHILD DEVELOPMENT

Refers to the physical, cognitive, emotional, and social growth that occurs throughout a child’s life. Child development is often seen as children being informed by the spheres of influence in their immediate and wider context including, for example, family, peers, school, neighbourhood, government services, and

media. For children today, the way they experience their digital environment is an important influence on their development.¹⁰⁷ (For detailed guidance on child development see Section 6.1.)

CHILDREN’S RIGHTS

The UNCRC¹⁰⁸ sets out a set of immutable rights to which all children are entitled. These uphold children’s right to be protected, to participate, and to life. General comment No. 25¹⁰⁹ describes how these rights apply in the digital environment.¹¹⁰

CHILDREN’S VOICE

Children have the right to be active participants in decisions that affect them.¹¹¹ This does not mean placing undue burden on children to raise or resolve problems (which is the responsibility of adults). Instead, adults must facilitate child-friendly ways to listen to children’s concerns, priorities, and preferred outcomes. Failure to do so is likely to lead to adult-led solutions that are inadequate.

VULNERABLE CHILDREN

Children are not a homogeneous group and some contexts (e.g., being a refugee or cared for child) or characteristics (e.g., race or gender) make them more or less vulnerable. When designing products, it is necessary to consider the vulnerabilities of all children likely to be impacted.¹¹²

B. AI¹¹³

ALGORITHM

A process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer.

ALGORITHMIC GROUPINGS

Services use algorithms to sort individuals, including children, into algorithmic groups¹¹⁴ that determine access to information, commercial, financial, educational, or employment opportunities; how goods are priced; how resources including public services are allocated; and the type and concentration of algorithmically recommended content.

ARTIFICIAL INTELLIGENCE (AI)

AI broadly refers to a range of algorithmic-based technologies and approaches that are aimed at mimicking human decision making.¹¹⁵

ARTIFICIAL INTELLIGENCE (AI) SYSTEM

See Section 1.1.

AI ECOSYSTEM

From the innovator’s perspective, an ‘AI ecosystem’ is a network of interconnected organisations (both commercial and non-commercial) and institutions (both governmental and non-governmental) that collectively enable the development, deployment, and management of AI.¹¹⁶ These actors can be AI enablers, producers, consumers, or regulators. *Enablers* provide the physical infrastructure and data management and processing abilities. *Producers* supply platform technologies or visualisation and analytics capabilities, which are used by *consumers* to build AI applications and use cases. *Regulators* are the legislators and regulatory bodies that set and enforce minimum standards. This ecosystem is interconnected by the data generated by consumers, which producers use to refine their algorithms, creating a feedback loop, and contributes to the unique data-driven ecosystem and consolidation of power among ‘tech giants’.

AI LIFECYCLE

The Organisation for Economic Co-operation and Development's (OECD) AI criteria define an AI system's lifecycle as involving: (i) "design, data and models", which is a context-dependent sequence encompassing planning and design, data collection and processing, as well as model building; (ii) "verification and validation", to test and evaluate the models; (iii) "deployment", making the system available for use; and (iv) "operation and monitoring". These phases often take place in an iterative manner and are not necessarily sequential. The decision to 'retire' an AI system may occur at any point during the operation and monitoring phase.¹¹⁷

AI SUPPLY CHAIN

Refers to a complex network of interconnected processes and data involved in the creation, development, deployment, and maintenance of AI models.¹¹⁸ This includes the sourcing of raw data, data preprocessing, algorithm development, model training, hardware provisioning, software development, deployment infrastructure, and ongoing monitoring and maintenance. Each stage involves different stakeholders,¹¹⁹ for example data providers, AI researchers, hardware manufacturers, software developers, and end users, ensuring the smooth delivery and operation of AI technologies. The supply chain may vary depending on the application and use cases.¹²⁰

ARTIFICIAL GENERAL INTELLIGENCE (AGI)

AGI technologies, also referred to as 'strong AI', are machines designed to perform a wide range of intelligent tasks, think abstractly, and adapt to new situations.¹²¹

ARTIFICIAL NARROW INTELLIGENCE (ANI)

ANI technologies, for example image and speech recognition systems, also called 'weak AI', are trained to perform specific tasks and operate within a predefined environment.¹²²

ASSUMPTIONS

Algorithmic assumptions are statements or conditions about the input, output, behaviour, performance, or correctness of an algorithm.¹²³

AUTOMATIC

A process or system that, under specific conditions, functions without human intervention.¹²⁴

BIAS

A disproportionate weight in favour of or against. Bias takes many forms. From a cognitive perspective, biases can be explicit or implicit/unconscious. Explicit biases are usually attitudes and beliefs that we are fully aware of, while unconscious biases are unintended, subtle, and subconscious, often learned through our past experiences. Unconscious biases can manifest as affinity bias, attribution bias, confirmation bias, halo effect, or horns effect.¹²⁵ In algorithmic systems, biases can be contributed by data, algorithm, or user biases. While these biases may come from different sources, depending on data sampling, rationales behind algorithmic designs or users' own prejudice, the human factors of these biases can often be attributed to the different forms of cognitive biases that are intrinsic to human judgement.¹²⁶

BUILDING

Refers to the process of identifying and collating data sets (inputs) and programming an algorithm that, when applied to data sets, will generate desired outputs.

CONTINUOUS LEARNING

The incremental training of an AI system that continues throughout the operational phase of the system's

lifecycle.¹²⁷ There are various risks that must be considered, especially for data that could be subject to rapid or unexpected shifts or drifts that could adversely impact the accuracy and performance of the AI system. Unstructured data, or a combination of structured and unstructured data, that process social or demographic data, may consequently pose risks of algorithmic bias and lurking discriminatory inferences.¹²⁸

FORMULA

A group of letters, numbers, or other symbols that represents a scientific or mathematical rule.

GENERAL PURPOSE AI SYSTEMS (FOUNDATION MODELS)¹²⁹

Artificial general intelligence (AGI) technologies with generative capabilities – referred to as 'general purpose AI', 'generative models' or 'foundation models' – are trained on a broad set of data that can be used for different tasks. These underlying models are made accessible to downstream developers through API (application programming interface) and open source access, and are used today as infrastructure by many companies to provide end users with downstream services.¹³⁰

GENERATIVE AI

A subset of artificial intelligence that uses generative models to produce text, images, videos, or other forms of data. These models often generate output in response to specific prompts. Generative AI systems learn the underlying patterns and structures of their training data, enabling them to create new data.¹³¹

INFERENCE

Reasoning by which conclusions are derived from known premises.¹³²

OPTIMISATION

To optimise an algorithm means to improve its performance so that it is more efficient and effective in solving a given problem.

SAMPLING

The process of selecting subsets of data from a larger data set intended to present patterns and trends similar to the larger data set for analysis.¹³³

SILENT SORTING

The process of assigning an algorithmic attribute or membership of an algorithmic group to individuals, including children, without their knowledge or consent.

TRAINING

AI model training is the process of feeding curated data to selected algorithms to help the system refine itself to produce accurate responses to queries.¹³⁴

C. Data¹³⁵

DATA DETERIORATION / DATA DECAY

This relates to the gradual deterioration of data quality over time. Tracking decline of data integrity and quality is essential to responsible governance.¹³⁶

DATA HYGIENE

Refers to the quality of the data on which the system is built. Data is examined for completeness, bias, and other factors that affect its usefulness for an AI system.¹³⁷

DATA LIFECYCLE

The sequence of stages that data goes through, from its initial generation or capture to its eventual archival and/or deletion at the end of its useful life.

DATA PROXIES

This refers to data points that are used as a substitute when exact data points are unavailable. Using proxies creates risks that inputs will not provide accurate outcomes.

DATA SECURITY

This refers to the protection of data used, generated, and stored by AI systems from unauthorised or unlawful access, processing, accidental loss, alteration, destruction or damage.¹³⁸

PERSONAL DATA

Information about an individual. Under data protection laws including the EU General Data Protection Regulation (GDPR) and UK data law, the person who the information is about needs to be identifiable, either directly or indirectly. So 'brown eyes' is not personal data about Sophie, but 'Sophie has brown eyes' or 'Everyone in this group has brown eyes and Sophie is in this group' are personal data. Sensitive personal data is personal data that a person is likely to consider as especially relevant to their identity, for example information about a person's race, ethnicity, gender, sexuality, health, political beliefs, immigration status, or financial circumstances.

Personal data includes inferred data, and it does not need to be accurate to meet the definition. For example, if an AI model infers that a child fits the criteria to be assigned to an algorithmic group of 'sad girls', this label is personal data irrespective of how the inference was made or whether it is true.¹³⁹

SCRAPING

Also known as data harvesting, this is the process of using automated systems to extract data from sources such as websites and social media.

VARIABLES

Data variables are characteristics in a dataset that can be measured, for example age, gender, and location.

D. Safety by design

AGE-APPROPRIATE DESIGN

Age-appropriate design anticipates the vulnerabilities, capacities, and needs of children at different developmental stages in the design of services and products.

ALGORITHMIC DESIGN

This is a design process that is based on algorithms, which are 'sets of mathematical Instructions or rules that... will help calculate an answer to a problem.'¹⁴⁰

'BY DESIGN AND DEFAULT'

A 'by design and default' approach requires developers to embed agreed or mandatory standards by design, for example standards for safety, privacy, or agency by design and by default. The Child Rights by Design framework provides guidance when considering children's best interests in the application and balancing of their protection and participatory rights under the UNCRC.¹⁴¹

CUMULATIVE HARM

The aggregate negative impacts that render harm at a societal level, which might not be significant individually, but can become substantial when accumulated over time.¹⁴²

IMPACT ASSESSMENT

This is an agreed process for anticipating, monitoring, addressing, and documenting the impact of service design and delivery on children.

MODERATION

A mechanism for resolving unanticipated and new risks that cannot be fully managed through the design of an AI system. It is most commonly used to surface and respond to potentially harmful or illegal content, contact, and conduct. Content may be surfaced through user reports or proactively by the AI system operator using a combination of human and automated moderation systems.¹⁴³

MONITORING

The continuous review process used to evaluate how AI systems are performing before and after launch. It is also used when considering the lifecycle of an AI system and making decisions on when to retire it.

OPPORTUNITY COST

This is where one route or course of action precludes the benefit of an alternative route or course of action.

RISK MITIGATION STRATEGY

A risk mitigation strategy provides the timeline and actions for eradicating risks identified and managing continual oversight to identify further risk as part of use.

SAFETY BY DESIGN

Addressing the known or anticipated risk of harm upstream through product design. The goal is to prevent, or substantially reduce, the risk of harm occurring in the first place.

USER INTERFACE

The means by which users interact with content to accomplish some goals.¹⁴⁴

6.4 Children and AI Design Code requirements checklist

1. Preparation	
PURPOSE	OUTCOMES
<p>The purpose of this stage is to ensure your operational and governance systems are sound.</p>	<p>When you have successfully completed the Preparation stage, you will have:</p> <ul style="list-style-type: none"> (a) Established a process for making decisions, including when and by whom. <input type="radio"/> (b) Created a project plan that conforms with the requirements of the Code. <input type="radio"/> (c) Provided a realistic estimate of resourcing needs (money, time, and people) that has been approved. <input type="radio"/> (d) Assembled a project team with the necessary skills, experience, and competencies. <input type="radio"/> (e) Assigned roles and responsibilities to team members and the Executive Leadership for all tasks. <input type="radio"/> (f) Made a written record of all the Preparation stage that has been reviewed and signed off (in writing) by the Executive Leadership. <input type="radio"/>

2. Intentions	
PURPOSE	OUTCOMES
<p>The purpose of this stage is to ensure you are clear on what you want your AI system to do and why.</p>	<p>When you have successfully completed the Intentions stage, you will have:</p> <ul style="list-style-type: none"> (a) Carried out an initial exploration of what you want your AI system to do and why (problem statement). <input type="radio"/> (b) Assessed your intentions against the criteria (Part Three) to identify and evaluate risk of non-conformity. <input type="radio"/> (c) Revised any aspects of your intentions that do not conform with the Code. <input type="radio"/> (d) Tested your revised intentions to ensure they now conform with the Code. <input type="radio"/> (e) Made a written record of your assessment process and the changes you have made in response that has been reviewed and signed off (in writing) by the Executive Leadership. <input type="radio"/> (f) Ensured your project plan aligns with your intentions. <input type="radio"/>

3. Data	
PURPOSE	OUTCOMES
<p>The purpose of this stage is to ensure the quality and appropriateness of the data used to build, train, and operate the AI system.</p>	<p>When you have successfully completed the Data stage, you will have:</p> <ul style="list-style-type: none"> (a) Carried out an audit of your proposed or existing data sources/inputs. <input type="radio"/> (b) Assessed your data inputs against the criteria to identify and evaluate the risk of non-conformity, including using appropriate testing if necessary. <input type="radio"/> (c) Revised any aspect of your data input that does not conform with the criteria. <input type="radio"/> (d) Tested your revised data input to ensure they now conform with the criteria. <input type="radio"/> (e) Made a written record of your assessment process and the changes you have made in response that has been reviewed and signed off (in writing) by the Executive Leadership. <input type="radio"/> (f) Provided in your project plan for ongoing monitoring of your data inputs, including ensuring that data generated by your AI system also conforms with the criteria. <input type="radio"/>

4. Development	
PURPOSE	OUTCOMES
<p>The purpose of this stage is to ensure that the way you design and train your AI system conforms with your Intentions Statement while meeting the criteria in the Code.</p>	<p>When you have successfully completed the Development stage, you will:</p> <ul style="list-style-type: none"> (a) Be clear on the instructions that will drive your AI system. <input type="radio"/> (b) Have assessed the instructions against the criteria to identify and evaluate risk of non-conformity using appropriate testing and consultation methods. <input type="radio"/> (c) Revised any aspect of your instructions that do not conform with the criteria. <input type="radio"/> (d) Tested your revised instructions to ensure they now conform with the criteria. <input type="radio"/> (e) Made a written record of your assessment process and the changes you have made in response. <input type="radio"/>

5. Deployment

PURPOSE	OUTCOMES
The purpose of this stage is to decide if your AI system is ready to be deployed.	When you have successfully completed the Deployment stage, you will:
	<ul style="list-style-type: none"> (a) Have completed all conformity assessments and testing. (b) Prepared a launch report for the Executive Leadership. (c) Conducted a launch review. (d) Received Executive Leadership approval or reverted to an earlier stage to address issues. (e) Made a written record of the launch review process. (f) Launched your AI system (if agreed).

6. Monitoring

PURPOSE	OUTCOMES
The purpose of this stage is to ensure you have systems and processes in place to monitor your AI system once it has been launched.	To meet your ongoing monitoring obligations, you will:
	<ul style="list-style-type: none"> (a) Have a plan and capacity for the continued monitoring of your AI system that has been approved by the Executive Leadership. (b) Have systems and processes to respond to issues identified through monitoring. (c) Run operational and team tests at regular intervals to ensure systems and processes continue to work effectively, and that personnel understand their roles and responsibilities. (d) Log monitoring outcomes, including incidents.



7. Transparency

PURPOSE	OUTCOMES
The purpose of this stage is to ensure you are transparent about your AI system and its impact on children or children in certain age groups.	At the end of this Transparency stage, you will have:
	<ul style="list-style-type: none"> (a) Developed a comprehensive transparency strategy (b) Assessed your data inputs against the criteria to identify and evaluate the risk of non-conformity, including using appropriate testing if necessary. (c) Developed all aspects of your transparency strategy collaboratively with relevant stakeholders, including children. (d) Taken account of the needs and capacities of children at different stages of development and those with additional vulnerabilities. (e) Identified ways in which you can provide users with key information about your AI system upfront and throughout the user journey, if your AI system is public facing. (f) Continually reviewed and updated your transparency strategy to ensure it is as user-friendly and useful as possible.

8. User reports and redress

PURPOSE	OUTCOMES
The purpose of this stage is to ensure you have effective report and redress systems and processes in place to facilitate user reports from or on behalf of children.	At the end of this stage, you will have:
	<ul style="list-style-type: none"> (a) A comprehensive user reporting strategy that takes account of the needs and capacities of children at different stages of development and those with additional vulnerabilities. (b) Created a way for parents, carers, and teachers to report on behalf of children that does not require being logged into or registered to your product or service. (c) Co-created your user reporting strategy with relevant stakeholders including children. (d) A protocol in place to inform regularly or in extremis the relevant authorities about emerging risks or incidents. (e) A plan to periodically review and update your user reporting strategy.

9. Retiring and moving on

 PURPOSE	 OUTCOMES
<p>The purpose of this stage is to assist you in planning to decommission your AI system, including predicting and monitoring its likely retirement date.</p>	<p>At the end of this stage, you will have:</p> <ul style="list-style-type: none"> (a) Agreed the criteria against which you will assess life expectancy and the cadence at which it will be reviewed. <input type="radio"/> (b) Carried out a preliminary review of your AI system's life expectancy. <input type="radio"/> (c) Conducted a decommissioning impact assessment for a planned and emergency retirement of your AI system. <input type="radio"/> (d) Been clear what steps you will need to take to retire your AI system, and the resources (people, time, and money) required to complete the process. <input type="radio"/> (e) Have emergency protocols in place in the event that it becomes necessary to retire your AI system at short notice. <input type="radio"/> (f) Secured written approval from the Executive Leadership of the retirement protocols, assessment, and planning. <input type="radio"/>

6.5 Snapshot case studies

This section contains a series of ‘snapshot’ case studies to illustrate how to apply the criteria at each stage of the Code. In reality, the process of following the Code will require a far more comprehensive assessment.

Snapshot on preparation

A company that operates a foundation model wants to review its current level of compliance with the Code criteria and is building a team to do so.

USEFUL QUESTIONS

What skills and competencies do your team members have?

- > The team comprises a project manager, an AI systems engineer specialising in safety, a data scientist, a privacy lawyer, and a member of the product policy team.

Who has overall responsibility for the review?

- > The review has been requested by the company’s Head of Compliance. They will lead the project and be responsible for deciding whether the model conforms with the Code and implementing changes (if needed).




PRELIMINARY VIEW

As currently formed, the team is missing some key skills and competencies. For example, it does not include expertise on child and adolescent development, on conducting research with impacted users (including children), and AI systems testing.

The governance and accountability provisions are insufficient because they do not include the Executive Leadership.

HOW DO I USE THE CRITERIA?

Here are some non-exhaustive examples of how the project team might evaluate the risks of the AI system against the criteria at the preparation stage (Section 5.1).

 ISSUE	 RISK	 POSSIBLE MITIGATION
The team is not comprehensive.	Without the necessary expertise, it isn’t possible to determine whether the AI system conforms with the Code criteria or what mitigations may be needed.	Using the guidance on team members’ roles and responsibilities (Section 6.2), assess the current team and identify gaps. Recruit additional members to address gaps from within the company or hire external experts.
The project is not being overseen by the Executive Leadership.	The project will not meet the accountability criteria.	Identify a member of the Executive Leadership to be the Senior Accountable Leader. Create a written policy on the governance and accountability systems and processes that will be followed throughout to ensure that the Executive Leadership are the final decision-makers for key decisions.
Planning may not be sufficiently detailed.	The project could fail if it isn’t set up properly.	Using the outcomes and guidance set out in Section 5.1, carry out a detailed review of all aspects of the project plan.

Snapshot on intentions

A government education department wants to build an AI system to automate pupil registration nationwide.

USEFUL QUESTIONS

What is the stated purpose?

- > To save teachers time by eliminating the pupil register each morning and afternoon.

How will the system work?

- > Each child will be given a fob that they must keep with them at all times. The fob will track their location using sensors placed around the school.

What information will be collected?

- > Whether a pupil is in school.
- > Whether they are in their class.
- > If they aren't in class, whether they are elsewhere on school premises.
- > Whether their absence from the classroom has been approved by their teacher.

How will this information be used?

- > To track and record attendance at school and in classes.
- > To support enforcement of school rules by identifying pupils who are out of class without permission.
- > To predict which pupils are high risk for poor attendance and disciplinary issues.
- > To inform planning for the school's special educational needs (SEN) and disciplinary intervention strategy.
- > To inform decisions to refer children to social services and other support services.
- > To predict academic attainment of individual pupils.

PRELIMINARY VIEW

The actual intentions go beyond the stated purpose.

HOW DO I USE THE CRITERIA?

Here are some non-exhaustive examples of how the project team might evaluate the risks of the AI system against the criteria at the intentions stage (Section 5.2).

ISSUE	RISK	POSSIBLE MITIGATION
The Intentions Statement is not comprehensive.	The disparity between stated and intended purpose means conformity with the criteria cannot be accurately assessed.	Rewrite the description of the intentions to reflect all intended uses before beginning the review process.
The intention may not conform with privacy criteria.	The level of monitoring of children's movements may disproportionately impact their right to privacy.	For all (remaining) intentions, consider whether the level of tracking is proportionate and if the intentions can be achieved in a less privacy-intrusive way.
The intention may not conform with fairness and reliability criteria.	It is unclear whether geolocation and attendance data are reliable predictors of academic and behavioural outcomes.	Take advice from domain experts on whether lesson and school attendance are reliable indicators of academic and behavioural outcomes. Consider removing intentions relating to academic and behavioural outcomes.

Snapshot on data

A local health authority wants to build an AI system to proactively share information about children undergoing treatment for mental health issues to enhance the care and support they receive.

USEFUL QUESTIONS

Who will the information be shared with?

- > Information will be shared between healthcare, social services and education professionals.

What information about children will the AI system be trained on?

- > Medical records of children currently undergoing treatment for mental illnesses.
- > Children’s services records of children currently undergoing treatment for mental illnesses (if applicable).
- > School records of children currently undergoing treatment for mental illnesses including family records, behaviour, visits to nurse, absence etc.

How will this information be collected?

- > The local authority will contract with a private company that will have access to the relevant databases. The company will then scan the records to create a project-specific data set.

Will the information be anonymised?

- > Yes, once it has been scanned and uploaded, the private company will anonymise it.

How will the security of the information be ensured?

- > The data set will be password protected, and the identity of anyone who accesses it will be recorded and time-logged.

PRELIMINARY VIEW

More careful consideration needs to be given as to how children’s data will be collected, processed, and secured.

HOW DO I USE THE CRITERIA?

Here are some non-exhaustive examples of how the project team might evaluate the risks of the AI system against the criteria at the data stage (Section 5.3).

ISSUE	RISK	POSSIBLE MITIGATION
The local authority plans to give API (application programming interface) access to highly sensitive personal data to a private company.	A security breach would expose children to significant harm.	Consult with security experts to understand the nature of the risk of a security breach via the API, and evaluate whether the current assurances about data security and privacy provided by the private company are sufficient.
The local authority has not considered what data records are needed.	The principle of data minimisation has not been applied.	The project team must establish what data is needed to build the AI system and limit data processing to what is strictly necessary.
Anonymisation of sensitive personal data will not take place until the data set has been compiled.	Failure to anonymise the data before it is shared compounds the possible negative impact of security breaches.	Adjust the project plan so that anonymisation takes place at the earliest possible stage.

Snapshot on development

A social media service wants to improve its social connections AI system to increase the number of follow/friend suggestions that users accept.

USEFUL QUESTIONS

What is the commercial objective of the project?

- > Increase data points on users to enhance profiling capabilities for advertising.
- > Increase the amount of time users spend on the platform by making their content feed more relevant.
- > Drive users to buy paid-for filters and effects.

What instructions will be used to direct the model?

- > Apply learning from user X's engagement metrics for content recommender systems.
- > Prioritise accounts that:
 - post content that is similar to content that user X watches when they spend 60 minutes or longer on platform;
 - post content that is similar to content that user X has engaged with (including negative engagement);
 - use paid-for filters and effects;
 - are most likely to accept a follow request.

Will accounts operated by children be included in the recommendation strategy?

Yes.

Were safety risks for children identified in the intentions and inputs stage?

Safety risks are being considered at the development stage.

If safety risks for children are identified, is it possible to apply additional protections to child users?

The level of certainty we have about the age of our users is currently low.

HOW DO I USE THE CRITERIA?

Here are some non-exhaustive examples of how the project team might evaluate the risks of the AI system against the criteria at the development stage (Section 5.4).

ISSUE	RISK	POSSIBLE MITIGATION
Safety risks should have been considered at earlier stages.	It may not be possible to address all risks at the development stage.	Revert back to the intentions stage.
The proposed use of children's accounts may be prohibited under the Digital Services Act.	The intended purpose may be unlawful in Europe.	Refer the proposal to legal experts for advice. Apply the standard of protection to children across all markets as best practice.
The recommendation logic includes tenuous social connections.	AI systems will recommend children to users they don't know.	Revise AI system instructions to include exceptions on follow recommendation logic for children or exclude children from the AI system.
If a child accepts a recommendation to follow an adult stranger, some safeguards on high-risk features such as private messaging fall away.	Adults will be able to message children they don't know in a private sphere.	Review the private messaging safeguards to avoid over-reliance on follow logic as a safeguard for child users.
Age assurance capabilities are low.	Even if we offer children heightened protections, we cannot guarantee that children will benefit.	Enhance age assurance capabilities or apply the highest standard of protection to all users.

Snapshot on deployment

A video-sharing platform wants to build an AI system to enhance its ads targeting capabilities for cosmetics, weight loss products, and cosmetic surgery.

USEFUL QUESTIONS

If deployed, which users would the AI system recommend the ads to?

- > Women and girls aged 13–30.
- > Women and girls who watch videos about depressing themes. Mental health and body image will be disproportionately targeted.

PRELIMINARY VIEW

In trying to optimise its ads targeting capabilities, the video-sharing platform appears to have built a model that has the ability to identify vulnerable girls and women. It has also captured children who are too young to consent to data processing.

HOW DO I USE THE CRITERIA?

Here are some non-exhaustive examples of how the project team might evaluate the risks of the AI system against the criteria at the deployment stage (Section 5.5).

ISSUE	RISK	POSSIBLE MITIGATION
The AI system will generate inferences about children’s mental health and wellbeing.	Children’s right to privacy will be negatively impacted.	Project as scoped is not viable.

Snapshot on monitoring

A leisure centre has deployed an AI system to automate oversight of its waterslides to save money on lifeguards. The AI system monitors the movements of swimmers as they queue, go down the slides, and clear the landing pool. This information is used to determine when it is safe for the next person to enter the slide and when to close the queue to manage risk of overcrowding.

 **USEFUL QUESTIONS**

Is the AI system already operational?

The AI system is being tested alongside current human-based oversight of the slides. While testing shows it is operating with a high level of accuracy and consistency, it will not be deployed until the monitoring plan has been approved by the Executive Leadership.

How long has the AI system been tested for?

The AI system was installed for testing in October and has subject to continued testing for four months.

What level of human oversight is envisaged for monitoring?

Lifeguards will be in place for the first 15 minutes following opening of the slides to monitor performance before authorising switch-over to the AI system. They will then repeat their review every three hours. The results of their review will be recorded. Lifeguards can override the AI system at any time and emergency protocol procedures set out the circumstances when lifeguards must initiate human override (e.g., if an accident is reported or one of the slides breaks).

What plans are in place to train your staff on the AI system?




To be confirmed.

 **PRELIMINARY VIEW**

At first glance, the leisure centre appears to have given thought to its monitoring strategy and has developed a proportionate human oversight plan. It must now think about training its staff to implement the protocols and also think further about its technical monitoring strategy.

 **HOW DO I USE THE CRITERIA?**

Here are some non-exhaustive examples of how the project team might evaluate the risks of the AI system against the criteria at the monitoring stage (Section 5.6).

 ISSUE	 RISK	 POSSIBLE MITIGATION
The AI system has only been tested during the winter months. The slides are much busier in the summer months.	It is not known whether the AI system will be as effective in managing queues if user numbers increase significantly.	Project team must consider additional post-deployment testing when user numbers go beyond a certain threshold with increased/full oversight during this time.
A plan to train staff on the system has not been developed.	The system relies on human oversight and may fail if those responsible for its operation do not know how to use it safely.	Create a training plan including compulsory in-person training for relevant personnel and a safety manual. Staffing rotas must reflect the need to have trained personnel available on every shift.

Snapshot on transparency

A tech company wants to launch a text-based generative model (large language model, LLM) for use by the general public. The model will help with tasks such as research and drafting. It is anticipated it will be used by children to help with schoolwork.

USEFUL QUESTIONS

What kinds of school/learning tasks can the LLM do?

Explain subjects (e.g., how to do equations or how to write an essay), create revision resources (e.g., quiz questions), correct and enhance drafting for essay questions, answer maths questions, comment on coursework skeleton plans, write coursework, and generate pseudo data sets for quantitative and qualitative research.

How accurate are the answers and advice that the LLM generates?

Accuracy rates are typically between 85–99%.

What plans do you have to inform students that the answers they receive may not be correct?

None.

Which of the tasks that the LLM can complete may contravene the school and exam board’s policies on cheating?

Unknown.

What steps do you have to inform students that using the LLM to complete schoolwork may constitute cheating and put them at risk of sanctions?

None.

What plans do you have to identify children and to surface information and warnings relating to use of the LLM for schoolwork?

None.

What is the universe of topics on which the LLM will return an answer? Will it answer more personal questions or questions unrelated to schoolwork?

The LLM can answer questions on almost any topic, but moderation systems are in place to prevent it generating content that is illegal.

PRELIMINARY VIEW

Children using the app may not understand that the answers it provides are inaccurate or that using it may expose them to disciplinary action for cheating. The issues raised go beyond transparency. For example, a full review of potential harms and the efficacy of the moderation strategy to mitigate risks is indicated. This snapshot focuses on transparency.

HOW DO I USE THE CRITERIA?

Here are some non-exhaustive examples of how the project team might evaluate the risks of the AI system against the criteria when thinking about transparency (Section 5.7).

ISSUE	RISK	POSSIBLE MITIGATION
Accuracy of information generated by the LLM is not explained.	Children may rely on inaccurate or incomplete information.	Improve the accuracy of the model. Include a visible warning about accuracy with all results. For results that appear to relate to schoolwork, include a warning that the LLM results may not be suitable for schoolwork and should be verified by a reliable source.
Implications of using the LLM to do schoolwork are not explained.	Children may not realise that using the LLM constitutes cheating and puts them at risk of punishment with long-term consequences for their academic future.	For results that appear to relate to schoolwork (e.g., essays, coursework, science/maths questions), include a warning that using the LLM to generate work that should be completed independently may constitute cheating and they should seek advice from their school.
The content moderation strategy is limited to illegal content.	Children may be exposed to content that is harmful but not illegal (e.g., suicide, self-harm, or graphic violence).	The project team must carry out a full review of all harms a child may be exposed to taking into account their age and intersectional vulnerabilities. If they choose not to do so or cannot guarantee the efficacy of the moderation strategy in preventing exposure to harm for children, they must age restrict access to some or all of the service. In addition to enforcing minimum age requirements, the service must provide clear and prominent information about its age and content policies, and take steps to bring both to the attention of children, parents or carers, and teachers.

Snapshot on user reports and redress

A tech company wants to launch an image-based generative model that enables users to access editing tools and effects to modify photographs and create new images.

USEFUL QUESTIONS

Is it possible for the AI system to modify or create an image of an identifiable child that could cause that child distress?

Yes.

What steps have you taken to mitigate this risk in earlier stages?

- > All filters are created by the company (not open sourced). They have been created using Safety by Design criteria to minimise the risk of misuse. For example, we do not offer filters that replicate plastic surgery, lighten skin tone by default, or which that users to generate images that are highly sexualised or show a person in embarrassing situations (e.g., going to the toilet).
- > We scan all images before they are presented to a user for harmful content and language.
- > Users who try to violate our policies are subject to warnings and other sanctions.
- > We monitor our performance to ensure our safety strategy is effective.

Have you put in place a child-friendly user complaints mechanism?

Yes.

Can people make a report if they aren't a registered user?

No.

PRELIMINARY VIEW

While the AI system controller has taken steps to ensure the system is safe and fair, they have not anticipated that children or adults acting on children's behalf who are not registered users may want to report an image that they have heard about or seen circulating on other platforms.

HOW DO I USE THE CRITERIA?

Here are some non-exhaustive examples of how the project team might evaluate the risks of the AI system against the criteria when thinking about user reports and redress (Section 5.8).

ISSUE	RISK	POSSIBLE MITIGATION
Only registered users can access the redress mechanisms.	The additional barrier of joining the platform to report may prevent children and those acting on their behalf from raising a legitimate report about the safety or privacy of a child.	Facilitate user reports via a website-based help centre that is accessible to everyone, and provide a helpline that is human-operated.

Snapshot on retirement

A health company uses an AI system that enables oncology consultants to generate personalised chemotherapy regimens for children. A breakthrough in treatment for leukaemia means that standards of care protocols for childhood leukaemia have been entirely revised.

USEFUL QUESTIONS

Are the recommendations from the AI system accurate and in line with the new care protocols?
No.
If the AI system was used to create a chemotherapy regime for a child and that regime was implemented, would that jeopardise the child's health and prospects for recovery?
Yes.

PRELIMINARY VIEW

While the AI system previously reflected best practice, it is now out of date and using it may put children at risk.

HOW DO I USE THE CRITERIA?

Here are some non-exhaustive examples of how the project team might evaluate the risks of the AI system against the criteria when thinking about retiring an AI system (Section 5.9).

ISSUE	RISK	POSSIBLE MITIGATION
The AI system is outdated.	Children will not receive optimum treatment and may have poorer outcomes.	Urgently retire the AI system and notify all parties in the supply chain so that it is not used by others.

Endnotes

- 1 ICO (Information Commissioner's Office) (2020) [Age appropriate design: A code of practice for online services](#).
- 2 European Commission (2022) [The Digital Services Act](#).
- 3 European Union (2024) [EU Artificial Intelligence Act](#).
- 4 Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media (2022) [Online Safety and Media Regulation Act 2022](#) [Ireland].
- 5 Republic of Singapore (2022) [Online Safety \(Miscellaneous Amendments\) Act 2022](#).
- 6 IMDA (Infocomm Media Development Authority) (2023) [Code of Practice for Online Safety](#) [Singapore].
- 7 Australian Government (2021) [Online Safety Act 2021](#).
- 8 For example, in 2023, leading technologists, including Elon Musk and Steve Wozniak, wrote an open letter calling for a pause in the development of the most powerful AI models, to develop and implement a set of shared safety protocols for advanced AI.
- 9 In the US, 7 out of 10 teens aged 13–18 said they had used at least one type of generative AI. While teens and their parents are equally likely to use search engines with AI-generated results (56% of teens and 55% of parents), teens are significantly more likely to have used: chatbots/text generators (51% of teens vs. 38% of parents), image generators (34% vs. 26%), and video generators (22% vs. 15%). See Madden, M., Calvin, A., Hasse, A., & Lenhart A. (2024) [The dawn of the AI era: Teens, parents, and the adoption of generative AI at home and school](#), Common Sense Media.
- 10 UNICEF (United Nations Children's Fund) (2023) [The state of the world's children: For every child, vaccination](#), UNICEF Innocenti – Global Office of Research and Foresight, April, p. 135.
- 11 The codification of those needs is the UN's Convention on the Rights of the Child (UNCRC). Widely reflected in legislation and culture, it touches on every area of public and private life across all corners of the globe, ensuring that children's development and rights are placed front and centre and made a societal concern. See OHCHR (Office of the High Commissioner for Human Rights) (1989) [Convention on the Rights of the Child](#), General Assembly resolution 44/25.
- 12 Mahomed, S., Aitken, M., Atabey, A., Wong, J., & Briggs, M. (2023) [AI, children's rights, & wellbeing: Transnational frameworks](#), The Alan Turing Institute, November.
- 13 European Union (2024) [EU Artificial Intelligence Act, Article 3: Definitions](#).
- 14 OECD (Organisation for Economic Co-operation and Development) (2019) [OECD AI Principles overview](#) [updated May 2024].
- 15 NIST (National Institute of Standards and Technology) (2023) [Artificial Intelligence Risk Management Framework \(AI RMF 1.0\)](#), US Department of Commerce, January.
- 16 See also Schmidhuber, J. (2014) [Who invented backpropagation?](#), AI Blog [updated 2020, 2022].
- 17 This list is informed by IEEE (2021) [Standard for an Age Appropriate Digital Services Framework based on the 5Rights Principles for Children](#), IEEE std 2089-2021, 30 November.
- 18 OHCHR (Office of the High Commissioner for Human Rights) (1989) [Convention on the Rights of the Child](#), General Assembly resolution 44/25, Article 1.
- 19 OHCHR (Office of the High Commissioner for Human Rights) (1989) [Convention on the Rights of the Child](#), General Assembly resolution 44/25.
- 20 OHCHR (Office of the High Commissioner for Human Rights) (2021) [General Comment no. 25 \(2021\) on children's rights in relation to the digital environment](#), 2 March.
- 21 See [Regulation \(EU\) 2024/1689 of the European Parliament and of the Council, laying down harmonised rules on artificial intelligence and amending Regulations \(EC\) No 300/2008, \(EU\) No 167/2013, \(EU\) No 168/2013, \(EU\) 2018/858, \(EU\) 2018/1139 and \(EU\) 2019/214](#).
- 22 The White House (no date) [Blueprint for an AI Bill of Rights: Making automated systems work for the American people](#).
- 23 See the White House's Executive Order 14110 on the [Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence](#).
- 24 Council of Europe (2024) [Framework Convention on artificial intelligence and human rights, democracy and the rule of law](#), Treaty Series No. 225.
- 25 Council of Europe (2024) [Methodology for the risk and impact of artificial intelligence systems from the point of view of human rights, democracy and the rule of law \(Huderia methodology\)](#), CAI(2024)16rev2.
- 26 NIST (National Institute of Standards and Technology) (2024) [Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile](#), NIST Trustworthy and Responsible AI, NIST AI 600-1.
- 27 Council of Europe (2018) [Guidelines to respect, protect and fulfil the rights of the child in the digital environment](#), Recommendation CM/Rec(2018)7 of the Committee of Ministers, September.
- 28 See the International Telecommunication Union's (ITU) [Guidelines on child online protection](#).
- 29 UNESCO (2024) [Recommendation on the ethics of artificial intelligence](#), 10 May.
- 30 UNICEF and Ministry for Foreign Affairs of Finland (2021) [Policy guidance on AI for children](#), November.
- 31 OHCHR (Office of the High Commissioner for Human Rights) (2021) [Artificial intelligence and privacy, and children's privacy – Report of the Special Rapporteur on the right to privacy](#), A/HRC/46/37, 25 January.
- 32 OECD (Organisation for Economic Co-operation and Development) (2021) [Recommendation of the Council on children in the digital environment](#) [adopted 16 February 2012, amended 31 May 2021].
- 33 OECD (Organisation for Economic Co-operation and Development) (2022) [Companion document to the OECD Recommendation on children in the digital environment](#), 20 May.
- 34 World Economic Forum (2022) [Artificial intelligence for children](#), Toolkit, March.
- 35 UN (United Nations) AI Advisory Body (2023) [Governing AI for humanity](#), Interim report, December.
- 36 G7 Hiroshima Summit (2023) [Hiroshima Process International guiding principles for organizations developing advanced AI systems](#).
- 37 AI Global Forum (2021) [Seoul Declaration for safe, innovative and inclusive AI by participants attending the Leaders' session of the AI Seoul Summit](#), Press release, 21 May.
- 38 See [Global Digital Compact](#).
- 39 OECD (Organisation for Economic Co-operation and Development) (2024) [Recommendation of the Council on artificial intelligence](#) [adopted 22 May 2019, amended 3 May 2024].
- 40 UN General Assembly (2024) [Draft resolution](#), A/78/L.49.
- 41 Council of Europe (2024) [Mapping study on the rights of the child and artificial intelligence: Legal frameworks that address AI in the context of children's rights](#), Prepared by the Alan Turing Institute and approved by the CDENF.
- 42 Mukherjee, S., Pothong, K., & Livingstone, S. (2021) [Child Rights Impact Assessment: A tool to realise child rights in the digital environment](#), 5Rights Foundation.
- 43 African Union (2024) [Continental Artificial Intelligence Strategy: Harnessing AI for Africa's Development and Prosperity](#).
- 44 AI Action Summit (2025) [Statement on Inclusive and Sustainable Artificial Intelligence for People and the Planet](#).
- 45 European Union (2024) [EU Artificial Intelligence Act, Article 25: Responsibilities along the AI value chain](#).
- 46 IEEE (2023) [Draft standard for the procurement of artificial intelligence and automated decision systems](#), IEEE P3119.
- 47 Quin, F., Weyns, D., Galster, M., & Costa Silva, C. (2024) [A/B testing: A systematic literature review](#), Journal of Systems and Software, 211, 112011.
- 48 Ganguli, D., Lovitt, L., Kernion, J., Askill, A., et al. (2022) [Red teaming language models to reduce harms: Methods, scaling behaviours, and lessons learned](#), arXiv, 2209.07858.
- 49 Nidhra, S. & Dondeti, J. (2012) [Black box and white box testing techniques: A literature review](#), International Journal of Embedded Systems and Applications (IJESA), 2(2), 29–50. doi: 10.5121/ijesa.2012.2204.
- 50 Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., et al. (2024) [International scientific report on the safety of advanced AI: Interim report](#), May.
- 51 See [Revealing Reality's What are avatars in research good for?](#)
- 52 Digital Futures Commission (2023) [Children's rights through children's eyes: A methodology for consulting children](#), March.
- 53 OHCHR (Office of the High Commissioner for Human Rights) (1989) [Convention on the Rights of the Child](#), General Assembly resolution 44/25.
- 54 Note that as of 2025 the United States has not yet ratified the UNCRC.

55 OHCHR (Office of the High Commissioner for Human Rights) (1989) [Convention on the Rights of the Child](#), General Assembly resolution 44/25.

56 OHCHR (Office of the High Commissioner for Human Rights) (2021) [General comment No. 25 on children's rights in relation to the digital environment](#), 2 March.

57 Informed by the Council of Europe's Gender Matters [guidance on intersectionality and multiple discrimination](#).

58 'Be private and secure' (US AI Bill of Rights). This means:
AI systems are designed and operated in accordance with general data principles (e.g., the General Data Protection Regulation, GDPR).
Children merit specific protection in relation to data privacy and design, and so AI systems must be operated in accordance with this principle. In the absence of local regulation, AI systems must operate in accordance with the standards prescribed in the UK's Age-Appropriate Design Code.
Children's data must be securely held by default, minimising the need for children to take proactive steps. Any aspect of a security system that requires proactive steps by children must be child-friendly etc.
Consent can only be relied on as the basis for processing children's data to operate an AI system where it can be meaningfully and appropriately given.
The domain in which the AI system operates and the impact on children's rights and opportunities must be considered. All domains that children use to activate their UNCRC rights are sensitive and require heightened protections. Health and education domains are highly sensitive.
Redress systems allow children to object to use of their data to build or run AI systems and to access information on what personal data is being processed etc.
Children are not tracked on a service or across other services to collect data to build or train AI systems, or to build a profile of children's characteristics, preferences, interests, behaviours, or beliefs.
Children's data (including inferred data) is not shared or sold to build AI systems.

59 ISO (2022) [International Standard Information technology – Artificial Intelligence concepts and terminology](#), ISO/IEC 22989, Chapter 5.15.

60 See Future-AI's [Traceability](#).

61 Livingstone, S., Cantwell, N., Özkul, D., Shekhawat, G., & Kidron, B. (2024) [The best interests of the child in the digital environment](#), Digital Futures for Children centre, London School of Economics and Political Science, and 5Rights Foundation.

62 While engineering standards such as ISO Guide 73: 2009 (now withdrawn) may use the term 'risk' to refer to an uncertain outcome that may be positive (an opportunity) or negative (a hazard), this Code adopts the common usage of risk to mean a negative outcome.

63 Examples of harmful or age-inappropriate content include promoting or normalising suicide, self-harm and eating disorders, mis- and disinformation, graphic violence, pornography, and bullying and harassment. Examples of illegal content and activity include grooming, child sexual abuse material, some forms of hate speech, and promoting terrorism.

64 See Internet Watch Foundation's (IWF) [Artificial intelligence \(AI\) and the production of child abuse imagery](#).

65 O'Hara, K. (2016) [The seven veils of privacy](#), IEEE Internet Computing, 20(2), 86–91.

66 Liu, C.-C., Liao, M.-G., Chang, C.-H., & Lin, H.-M. (2022) [An analysis of children's interaction with an AI chatbot and its impact on their interest in reading](#), Computers & Education, 189, 104576.

67 See Revealing Reality's [Your new best friend: Generative AI](#).

68 eSafety Commissioner (2025) [AI chatbots and companions – Risks to children and young people](#), Online safety advisory, 18 February. See also Ofcom (2024) [Open letter to UK online service providers regarding Generative AI and chatbots](#), 8 November.

69 This process is achieved to a large extent through interactions with peers, which can be turbulent and unpredictable and require an element of compromise. If children find synthetic relationships more straightforward and rewarding in the short term, they may not gain sufficient life experience to forge successful platonic, romantic, and professional relationships in the long term.

70 Article 9 of the [Artificial Intelligence Act](#) defines a risk management system 'as a continuous iterative process planned and run throughout the entire lifecycle of a high-risk AI system, requiring regular systematic review and updating. It shall comprise the following steps:

(a) the identification and analysis of the known and the reasonably foreseeable risks that the high-risk AI system can pose to health, safety or fundamental rights when the high-risk AI system is used in accordance with its intended purpose.

(b) the estimation and evaluation of the risks that may emerge when the high-risk AI system is used in accordance with its intended purpose, and under conditions of reasonably foreseeable misuse.

(c) the evaluation of other risks possibly arising, based on the analysis of data gathered from the post-market monitoring system referred to in Article 72.

(d) the adoption of appropriate and targeted risk management measures designed to address the risks identified pursuant to point (a).'

71 See AI Ethics and Governance in Practice's [AI accountability in practice](#). This must take the form of a process log or similar documentation technique that ensures end-to-end traceability, answerability, and auditability.

72 European Union (2024) [EU Artificial Intelligence Act, Article 10: Data and data governance](#).

73 European Union (2024) [EU Artificial Intelligence Act, Article 14: Human oversight](#).

74 Sometimes referred to as OKR (objectives and key results).

75 Based on and adapted from IEEE (2021) [Standard for an Age Appropriate Digital Services Framework based on the 5Rights Principles for Children](#), IEEE std 2089-2021, 30 November.

76 See Livingstone, S. (2024) [The best interests of the child in the digital environment](#), Digital Futures for Children centre, March.

77 Based on 5Rights Foundation (2022) [Shedding light on AI: A framework for algorithmic oversight](#), June.

78 5Rights Foundation (2022) [Shedding light on AI: A framework for algorithmic oversight](#), June.

79 ISO (2022) [International Standard Information technology – Artificial Intelligence concepts and terminology](#), ISO/IEC 22989.

80 See the IEEE's [Online age verification certification program](#).

81 This might work in the same way as the UK government's '[Yellow Card](#)' scheme for reporting adverse incidents with medicines and medical devices.

82 Council of Europe (2024) [Methodology for the risk and impact of artificial intelligence systems from the point of view of human rights, democracy and the rule of law \(Huderia methodology\)](#), CAI(2024)16rev2.

83 European Union (2024) [EU Artificial Intelligence Act, Article 13: Transparency and provision of information to deployers](#).

84 Alan Turing Institute (2025) [The children's manifesto for the future of AI: Making our voices heard](#), February.

85 ISO (2022) [International Standard Information technology – Artificial Intelligence concepts and terminology](#), ISO/IEC 22989, p. 31.

86 Wood, S. (2024) [Impact of regulation on children's digital lives](#), Digital Futures for Children centre, London School of Economics and Political Science, and 5Rights Foundation, May.

87 See Google's AI [Model Card Toolkit](#) for guidance on a standardised approach.

88 For a recent example, see [Meta and Center for Open Science Open request for proposals for research on social media and youth well-being using Instagram data](#).

89 For further resources on AI auditing, see, for example, OECD's AI Policy Observatory, [AI Auditing](#) and [AI Standards Database](#).

90 For further resources on standards, see, for example, [Catalogue of tools & metrics for trustworthy AI](#).

91 See UKAS '[About us](#)'.

92 See CEN and CENELEC on '[European standardization](#)'.

93 See, for example, the [ICO certification schemes](#).

94 Stevens, E. (2023) [What are user research ethics? The 5 most important ethical considerations in UX research](#), UX Design Institute Blog, 28 June.

95 See the National Institute of Standards and Technology's (NIST) [Trustworthy & Responsible AI Resource Center](#) on '[Govern](#)'.

96 Including research carried out by Professor Sonia Livingstone at the London School of Economics and Political Science – ICO (Information Commissioner's Office) (2018) [Call for evidence: Age Appropriate Design Code; Annex B of the ICO's Age-Appropriate Design Code](#); 5Rights Foundation's [Disrupted Childhood reports \(2017 and 2023\)](#); and Ofcom's research on [Child development ages, stages, and online behaviour](#) (2024).

- 97 Green, L., Haddon, L., Livingstone, S., O'Neill, B., Stevenson, K., & Holloway, D. (2024) Digital media use in early childhood: Birth to six, Bloomsbury.
- 98 Auxier, B., Anderson, M., Perrin, A., & Turner, E., (2020), Children's engagement with digital devices, screen time, Pew Research Centre.
- 99 Ofcom., (2024), Children and parents: media use and attitudes report 2024.
- 100 The United States Code (formally the Code of Laws of the United States of America) is the official codification of the general and permanent federal statutes of the United States.
- 101 Title 42 s.675(11)(A)(i) of the United States Code.
- 102 5Rights Foundation (2021) But how do they know it is a child? Age assurance in the digital world, October.
- 103 For example, Article 19 of the UNCRC establishes a child's right to be protected from physical or mental violence, injury or abuse, neglect or negligent treatment, and maltreatment or exploitation, including sexual abuse. See OHCHR (Office of the High Commissioner for Human Rights) (1989) Convention on the Rights of the Child, General Assembly resolution 44/25.
- 104 For example, Article 14 of the UNCRC establishes a child's right to freedom of thought, conscience, and religion. OHCHR (Office of the High Commissioner for Human Rights) (1989) Convention on the Rights of the Child, General Assembly resolution 44/25.
- 105 'In all actions concerning children, whether undertaken by public or private social welfare institutions, courts of law, administrative authorities, or legislative bodies, the best interests of the child shall be a primary consideration' (Article 3, UNCRC). OHCHR (Office of the High Commissioner for Human Rights) (1989) Convention on the Rights of the Child, General Assembly resolution 44/25.
- 106 OHCHR (Office of the High Commissioner for Human Rights) (2021) General comment No. 25 (2021) on children's rights in relation to the digital environment, 2 March, paragraph 12; see also Livingstone, S., Cantwell, N., Özkul, D., Shekhawat, G., & Kidron, B. (2024) The best interests of the child in the digital environment, Digital Futures for Children centre, London School of Economics and Political Science, and 5Rights Foundation.
- 107 Smahel, D., Machakova, H., Mascheroni, G., Dedkova, L., et al. (2020) EU Kids Online 2020: Survey results from 19 countries, EU Kids Online, doi: 10.21953/lse.47fdeqj01ofo.
- 108 OHCHR (Office of the High Commissioner for Human Rights) (1989) Convention on the Rights of the Child, General Assembly resolution 44/25.
- 109 OHCHR (Office of the High Commissioner for Human Rights) (2021) General comment No. 25 (2021) on children's rights in relation to the digital environment, 2 March.
- 110 Child Rights by Design sets out 11 Principles against which the digital environment must be designed to ensure children's UNCRC rights are upheld. See Livingstone, S. & Pothong, K. (2023) Child Rights by Design: Guidance for innovators of digital products and services used by children, Digital Futures Commission and 5Rights Foundation.
- 111 'States Parties shall assure to the child who is capable of forming his or her own views the right to express those views freely in all matters affecting the child, the views of the child being given due weight in accordance with the age and maturity of the child' (Article 12, UNCRC). See OHCHR (Office of the High Commissioner for Human Rights) (1989) Convention on the Rights of the Child, General Assembly resolution 44/25.
- 112 OHCHR (Office of the High Commissioner for Human Rights) (2021) General comment No. 25 (2021) on children's rights in relation to the digital environment, 2 March, paragraph 11.
- 113 Jones, E., (2023) What is a foundation model?, Ada Lovelace Institute, Resource, 17 July; and ISO (2022) International Standard Information technology – Artificial Intelligence concepts and terminology, ISO/IEC 22989.
- 114 Wachter, S. (2022) The theory of artificial immutability: Protecting algorithmic groups under anti-discrimination law, Tulane Law Review, 97, 149.
- 115 Guidance published by the UK's Information Commissioner's Office (ICO) and the Alan Turing Institute in May 2020: Explaining decisions made with AI.
- 116 Jacobides, M.G., Brusoni, S., & Candelon, F. (2021) The evolutionary dynamics of the artificial intelligence ecosystem, Strategy Science, 6(4), 412–435.
- 117 The OECD's AI Principles overview (updated in May 2024); ISO (2022) International Standard Information technology – Artificial Intelligence concepts and terminology, ISO/IEC 22989, p. 35 and Annex A.; and IEEE SA (Standards Authority) (2017) ISO/IEC/IEEE International Standard – Systems and software engineering -- Software life cycle processes, IEEE/ISO/IEC 12207-2017.
- 118 Brown, I. (2023) Allocating accountability in AI supply chains, Ada Lovelace Institute, Resource, 29 June.
- 119 Definitions of key stakeholders in an AI supply chain can be found in Article 3 of the EU Artificial Intelligence Act.
- 120 Brown, I. (2023) Allocating accountability in AI supply chains, Ada Lovelace Institute, Resource, 29 June.
- 121 Madiega, T. (2023) General-purpose artificial intelligence, Briefing, European Parliamentary Research Service, March.
- 122 Madiega, T. (2023) General-purpose artificial intelligence, Briefing, European Parliamentary Research Service, March.
- 123 How do you check your algorithm assumptions? generated by LinkedIn's AI tool [accessed 18 February 2025].
- 124 ISO and IEC (2022) Information technology – Artificial intelligence – Artificial intelligence concepts and terminology, ISO/IEC 22989.
- 125 Houser, K.A. (2019) Can AI solve the diversity problem in the tech industry? Mitigating noise and bias in employment decision-making, Stanford Technology Law Review, 22, 290.
- 126 Ferrara, E. (2024) Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies, Sci, 6(1), 3.
- 127 ISO and IEC (2022) Information technology – Artificial intelligence – Artificial intelligence concepts and terminology, ISO/IEC 22989.
- 128 Council of Europe (2024) Methodology for the risk and impact of artificial intelligence systems from the point of view of human rights, democracy and the rule of law (Huderia methodology), CAI(2024)16rev2.
- 129 Triguero, I., Molina, D., Poyatos, J., Del Ser, J., & Herrera, F. (2024) General Purpose Artificial Intelligence Systems (GPAIS): Properties, definition, taxonomy, societal implications and res-ponsible governance, Information Fusion, 103, 102135.
- 130 Madiega, T. (2023) General-purpose artificial intelligence, Briefing, European Parliamentary Research Service, March.
- 131 Zewe, A. (2023) Explained: Generative AI, MIT News, 9 November; see also the Wikipedia page on Generative artificial intelligence.
- 132 ISO (2022) International Standard Information technology – Artificial Intelligence concepts and terminology, ISO/IEC 22989.
- 133 ISO (2022) International Standard Information technology – Artificial Intelligence concepts and terminology, ISO/IEC 22989.
- 134 Chen, M. (2023) What is AI model training & why is it important?, Oracle United Kingdom, 6 December.
- 135 Jones, E. (2023) What is a foundation model?, Ada Lovelace Institute, Resource, 17 July.
- 136 ISO (2022) International Standard Information technology – Artificial Intelligence concepts and terminology, ISO/IEC 22989.
- 137 ISO (2022) International Standard Information technology – Artificial Intelligence concepts and terminology, ISO/IEC 22989.
- 138 See ICO guidance, How should we assess security and data minimisation in AI?
- 139 See ICO guidance, Can opinions or inferences about people be personal data?
- 140 Caetano, I., Santos, L., & Leitão, A. (2020) Computational design in architecture: Defining parametric, generative, and algorithmic design, Frontiers of Architectural Research, 9(2), 287–300.
- 141 Livingstone, S. & Pothong, K. (2023) Child Rights by Design: Guidance for innovators of digital products and services used by children, Digital Futures Commission, 5Rights Foundation.
- 142 Janjeva, A., Mulani, N., Powell, R., Whittlestone, J., & Avin, S. (2023) Strengthening resilience to AI risk: A guide for UK policymakers, Centre for Emerging Technology and Security Briefing Paper, August.
- 143 For more information, see the Trust & Safety Professional Association's (TSPA) Content moderation and operations.
- 144 Blair-Early, A. & Zender, M. (2008) User interface design principles for interaction design, Design Issues, 24(3), 85–107.

