Education Services Australia

AI in Australian Education Snapshot:

Principles, Policy, and Practice

August 2023

Contents

Acknowledgments

Author: Daniel Ingvarson
International education and technology consultant. Advisor to education systems, technology companies on Googles internal Generative AI advisory for kids and family, education measurement and data analytics. Dan@the.inter.net.au

Contributing author: Beth Havinga

We would like to thank the many people who have contributed to the creation of the paper through their time, interviews and surveys. Including, Matt Deeble, Jeremy Rochelle, Joseph South, Keith Kruger, Juliette Norrmen-smith, Alex Voß, Leon Furze, Laurie Forcier, Steve Midgley, Doug Jaffe, Lukman Ramsey, Jim Larimore, Eric Nrntrup, Erin Mote, Sandra Milligan, Lauren Sayer, Jim Knight, Rose Luckin.

Originally published: August 2023

Cover image generated using Adobe Firefly AI

1. Executive summary

Governments, education systems, and non-profit organisations around the world are responding to the challenges and opportunities presented by generative AI, and Large Language Models (LLMs) in particular. Responses are typically led by the creation of indices of principles that provide context and objectives to guide implementation. This report has found just over 300 sets of such principles with only a small number specific to education.

There is currently a lot of activity related to AI principles, and this report's second section is an initial review of that activity. There is clear convergence between governments' concerns and the sets of AI principles which are being put forward for consideration. An analysis of the public consultation draft version of the Australian Education AI Taskforce's principles shows them to be in line with other frameworks and we do not offer a critique of the National AI in Schools Framework directly.

In relation to moving from 'principles into practice', this review established that governments from across the globe are creating laudable sets of principles that are highly similar and which do address the key issues, but which are difficult to enact as an increasing range of market options for implementing AI emerge. Enacting some of the principles may place substantial new burdens on schools to amend processes and monitor compliance. Others would require new agreements in relation to data management and protection, and others would require new measurements of accountability and transparency with vendors, schools, and systems. A useful approach to manage this complexity is to triage principles into three categories: those that can be implemented now with a focus on features which support a level of 'base safety'; those that clearly require new changes in rules or agreements; and those that need further research or further definition to be feasible.

The first section of this report explains how important LLM developments demonstrate that we are all at the start of an ongoing development process for AI. In this regard our responses (principles, policies, practices etc) need to be designed and implemented in a way that evolves as these tools, our ability to measure them, industry's ability to implement them and our use of them, change over time.

In our experience, the market pressure to rapidly evolve AI products is greater than any previous technological change. This has encouraged the development of a multi-tiered system of AI's which is just one illustration of the complicated AI landscape. The current main players are creating new capabilities with huge models (which we name as Large Foundational Models (or LFMs)). These are likely to be restricted down to a dozen or so major companies because of cost, data availability and computing power required. Another tier of AI players (in the hundreds) is creating smaller specific models using open-source or pared down commercial models. There is a further tier who are re-using all types of Gen AI within their products, many of them are creating local customisations, changing the behaviour of the Gen AI for their local context. This is likely to grow to tens of thousands of companies and could lead to a proliferation. The developments in each tier impact how systems and schools can enact AI policy and practice in ways that are consistent with the relevant principles.

International governments' responses fall into policy model categories of: AI stand-alone policy; AI integrated per sector; and a thematic approach. At the time of writing the EU's laws are the only example of an external enforcement mechanism being implemented to date, and it will take time to see how industry responds and if the multinational risk-based approach, which requires inclusion in the law of pre-identified of areas of high risk,[1] will be sustainable.

The non-government and not-for-profit sector has moved at speed to create resources for schools, for learning about and with, AI. There are resources for each level of schooling to assist with policy development, guidelines in schools, PD support for teachers, and resources to use in classrooms. Many commercial offerings take the form of free courses and materials to help schools use AI and many warrant inclusion in schools' AI toolkits with quality and practical advice, mainly aimed at the classroom teachers.

The organisations focussed on school or educator activity are working from the 'bottom up', whilst the government is setting a vision through principles that is 'top down'. Presently there is a gap with local solutions lacking the context of, or express alignment to, the principles. Conversely the national frameworks are often not considering how their principles can be implemented. We recommend a pragmatic approach that prioritises those elements of a principle that can be enacted now.

An example of this pragmatism is the voluntary commitments from the large tech firms recently secured by the Biden Administration, which is a significantly divergent approach to the EU. Instead of starting with the identified set of principles, the approach creates a compromise for action. They appear to have struck a balance between what is technically achievable and the public's main concerns. They set objectives of "secure, safe and trustworthy" that are defined in a technically achievable way with the tools within reach in an acceptable timeframe.

The gap identified is, therefore, not between one or another of the sets of principles created across the world, but rather between moving those principles into practice. There is a need for a multi-step approach where the different needs of the participants are considered and catered for at a sector level, school level, a student level as well as with the public and the different players of the industry ecosystem in mind. This approach needs to be addressed within the agreed framework of the principles and enacted as the capabilities make this feasible.

The EdSAFE AI alliance and the Australian organisation IAMAI have each created policy guidance frameworks to support the refinement of the principles to move them into practice. They begin by assessing each principle's definition against criteria of clarity, measurability, enforceability and urgency. This guides actions that balance what is needed and what is possible now.

We recommend an approach that has activities working in different time horizons. The immediate time horizon, like the US approach, is an early response to create a level of 'base

1 The main area of high risk identified for education is negative impact of education opportunity from inaccurate grading by AI.

safety' that schools are able to enact without significant burden and which is technically feasible. The immediate needs of 'base safety' for education include action on data protection, audit and accountability, benchmarks and support for schools' main concerns around academic integrity, staff development and 'humans in the loop' policies to ensure AI is not directly impacting students without teachers' knowledge.

The next time horizon deals with research and support for evidence-based best practice. What measurement can school leadership use to know how much use of AI is too much for administrators, for teachers to create resources or for student use in learning? Recent investments in the creation of high-quality teaching resources should be reviewed with consideration given to these new AI capabilities. Importantly, the potential of AI for effectiveness should be addressed with a special focus. The current focus on de-risking is important, but investment is urgently needed in research focused how AI can change teaching and learning, save teachers' time, address equity gaps, modernise curriculum, expand our assessment options, and support a wider range of students' needs while improving outcomes for all.

The final time horizon works on structural questions, such as jurisdictional governance and procurement from the foreign-owned and operated Large Foundational Models. This is where Australia might look to develop a mandated EU-style approach for important items, cognisant of a global market, maturing AI capabilities and the development of new measures for ethical and bias based appraisals. We will have to consider how to integrate local cultural expectations, alignment to values and the balancing of the different views of truth and facts and what makes up our curriculum, in an environment of hundreds or thousands of AI enabled software platforms with different implementation models.

We need rapid progress on all three horizons, even if their resolution and impact on school choices differ in time. Thinking in this way will enable us to move from principles into practice and ensure the principles are evolving along with the AI developments.

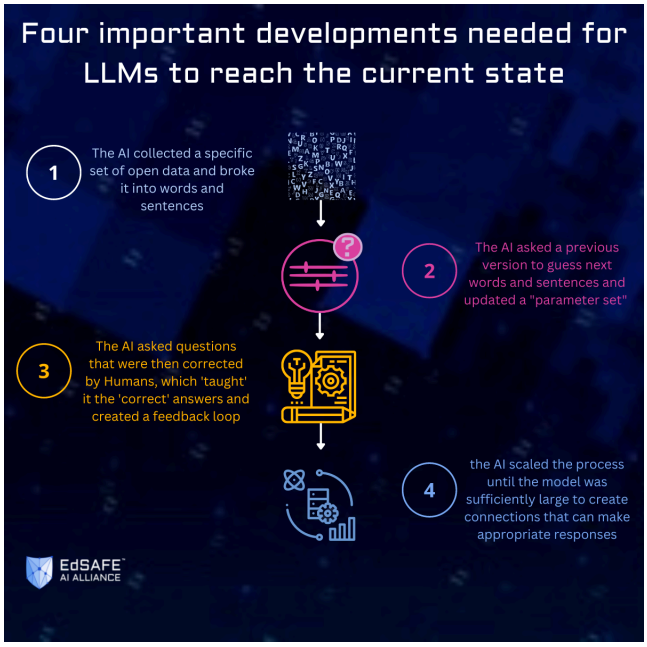
2. Introduction to LLMs

The following section provides a non-technical introduction to a large language model and key terms and concepts which are integral to this report.

What is a large language model?

A large language model (LLM) is a type of machine learning model that can perform a variety of natural language processing (NLP) tasks. This is what gives these types of models

their ‘human-like’ interaction capacities as they can generate, translate, and classify text, and answer in a conversational manner.



“Large” in this case is a reference to the number of parameters that the model can change autonomously as it learns. Some of the current LLMs have up to hundreds of billions of parameters. In order to reach this number of parameters, LLMs require immense amounts of data so that they can be trained. The training creates the relationships between words by the LLM predicting the next word within a sentence. This method of checking predictions and getting feedback helps to hone the model until it can deliver an appropriate level of accuracy. This accuracy reflects the creation of a conceptual network or map of concepts that exist in our world, behind the words, into their conceptual relationships.

The systems then go through an extensive refinement stage where humans (often low paid and in developing counties) correct answers. This is called Reinforcement Learning Human Feedback (RLHF) and for GTP 4 it took 8 months. It is the critical refinement stage and sets apart models. Once an LLM has been RLHF trained, it can be fine-tuned for a wide range of tasks, including:

- Building conversational chatbots like ChatGPT.
- Generating text for product descriptions, blog posts and articles.
- Answering frequently asked questions (FAQs) and routing customer inquiries to the most appropriate human.
- Analysing feedback from email, social media posts and product reviews.
- Translating business content into different languages.
- Classifying and categorising large amounts of text data for more efficient processing and analysis.

Generative AI text training

Having accessed and been trained on a lot of the available datasets in the world,[2] the current language models are completing an establishment phase where, like us, they have read and learned. These systems do not maintain a copy of that original data (only the language relationships which impact their existing models) of which any one piece of content provides an almost undetectably small amount of learning. This has some important policy considerations;

[2] https://www.theatlantic.com/technology/archive/2023/01/artificial-intelligence-ai-chatgpt-dall-e-2-learning/672754/

- The results of the systems are, **in some cases**, finding diminishing returns on additional data. Having already learned from much of the open and quality data on the internet. New sources and tactics for training data currently in development may be more intrusive or personal and require a policy trade-off and opening of currently closed sources.
- A regulation approach, which can inform the consent practices required as systems need to learn from us should be identified. The current terms of use is a now familiar approach needs review in this light, where data is inputted and then available for world-wide reuse by the platform provider with the learnings belonging to the platform.
- Because the training of LLMs has, in many cases, already taken place or is currently undergoing, any proposed regulations to control the data training being used to train generative AI LLMs will only affect subsequent models and potentially not have any effect at all.
- It is vital to evaluate what policy or regulation work is **possible** to ensure that there is safety, efficacy and equity in the use of these models in education. Especially as 'jailbreak' hacks are already being found,[3] which circumvent existing safety measures.

Probabilistic networks are evolving into foundational networks

Conceptual networks are made by trying to guess the next word. And if the system guesses correctly, it updates the network. If it gets it wrong, the system updates the network accordingly. It does that for every web page, every Wikipedia article, every science article, every Reddit article etc. This includes languages, where the concept of 'hello' is related to that of 'Gday' which can then easily be extended to 'Bonjour', then all the concepts and relationships already related to 'hello' can then be reused in another language.

The key discovery in 2022 was that larger probabilistic networks (called models) then successfully perform tasks they were **not** trained on. I.e., they exhibit unexpected results which are called 'emergent' abilities. However, until we find that ability by asking questions, it is not known; it is in the software but hidden until found. That is, this is the first software which ships with an unknown set of capabilities, and the bigger and more complex the model, the more emergent capabilities lay dormant waiting to be discovered.

There are also probability relationships in images, videos, and audio, where the same base concepts and their relationships are reapplied to these new data sets. That is, it turns out the same make-a-guess and check and update the concept network applies to words can be applied to a range of other circumstances. This is called multimodal, and enables the concepts expressed in an image (such as a smile and a wave) for hello to be related to hello in language much like our brain does. However, it goes further, with movement being able to be both learnt and used from the same Foundational Network. 'Hello' for instance is a concept that can expressed in movement by a wave. This appears to be analogous to the learning a child does by falling down or the learning from success/failure from movement, based on the same correct/incorrect/update probability of success formula. This approach

[3] https://futurism.com/hack-deranged-alter-ego-chatgpt

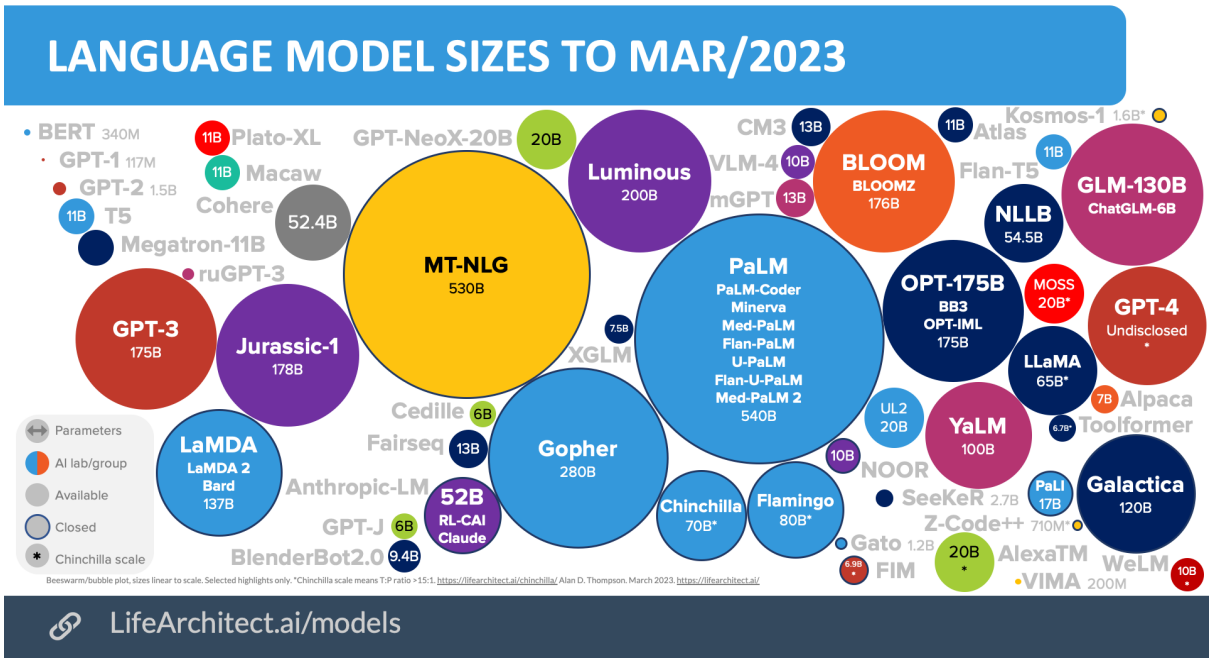
has proven so effective that it outperformed 20 years the robotics company Boston Dynamics spending over $1 billion in research and development), was overtaken by the likes of X1[4] in less than a year at fraction the of the spend.

It initially took 15 years to get an AI to be at the level of human performance in certain categories. It now takes only a few months. The speed of change is astronomical and the next target is artificial general intelligence (AGI) which OpenAI defines as 'highly autonomous systems that outperform humans at most economically valuable work,[5] and Dr Alan Tompson (leader in AI analysis) predicts will arrive in 2025.[6] The below diagram shows the number of models now in development, showing we are at the very start of this change. Those with a societal wide view liken the change to the industrial revolution.[7] Within a few short years. ChatGPT will look like a novel historical tool, a bit like pulling out an old cell phone.

It will profoundly impact industry with announcements already from IBM to cut 30% jobs[8], BT in UK to cut 55,000 jobs[9] in the future. However it's also happening now with Fortune reporting that 4,000 jobs were lost to AI in May 2023.[10] The issue in education is not the reduction in workforce (no one we have talked to is talking about increasing class size significantly) but how to use the predicted "AI Dividend"[11] to create a more safe & effective education system while decreasing the workload and improving the attractiveness of teaching to encourage more and better people to join the profession. This requires a vision for education not just a vision for managing or de-risking AI.[12]

4 https://www.1x.tech/neo

5 https://openai.com/charter
6 https://lifearchitect.substack.com/p/the-memo-17aug2023
7 https://www.realclearpolitics.com/video/2023/05/06/david_brooks_ai_is_the_industrial_revolution_it_will_have_pervasive_effects_on_society_and_culture.html
8 https://www.dailymail.co.uk/news/article-12036965/IBM-says-pause-hiring-CEO-says-7-800-non-customer-facing-roles-replaced.html
9 https://www.theguardian.com/business/2023/may/18/bt-cut-jobs-telecoms-group-workforce
10 https://fortune.com/2023/06/02/ai-job-cuts-layoffs-tech-industry-challenger-grey/
11 Concept from D.Ingvarson elaborated in various presentations including https://www.youtube.com/watch?v=WyC3bluq6k0
12 D.Ingvarson's presentation to education ministerial conference in London 03/20203 outlines solving educations intractable and long term issues as the goal for AI.



There are [indexes of what was included in models](13) before March 15 of this year. As of this date, however, Open AI closed access[14] to the information regarding its training data setting off the closed AI phase where the large players no longer provide information on their models. It. It is important to have transparency regarding how systems were developed and what kind of information they were trained on, because the underlying conceptual relationships will persist even though we use the post data training RLHF to retrain the models to address the values alignment problem. This serves to mask what is in the base model but not remove it (see jailbreaking[15]). This an expensive process which took eight months on GPT4 and is current our primary 'safety' layer.

[13] https://docs.google.com/spreadsheets/d/1O5KVQW1Hx5ZAkcg8AIRjbQLQzx2wVaLl0SqUu-ir9Fs/edit
[14] https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview
[15] https://www.techopedia.com/what-is-jailbreaking-in-ai-models-like-chatgpt

An additional approach championed by Anthropic is 'constitutional AI[16]' where one AI, which has already been through a data training process, is given a rule book (or constitution) to train another AI. Each of these processes is imperfect, however they are currently our key method for guiding a model in relation to our values system.

This issue has implications for the use of AI models in education, as there is a link between what data is input, and what RLHF retraining is done, by whom and which facts[17] are selected as truth to reinforce. and then what in education we deem as correct, and even further, what is effective in education, what research we are endorsing[18] and what is aligned with our culture and beliefs.

The Whitehouse science and technology director, Arari Prabhakar, stated on July 21st 2023:

"We do not have tools or methods today to know when an AI model is safe and effective... That is, we don't know how to tell if they are safe"

This appears to support the notion stated above, that new tools and mechanisms are required in order to monitor or understand LLMs, and it is important for education systems to develop a path to safe and effective AI responses from LLMs.

There are a number of important issues on this path. Some issues such as an ability to audit LLMs chat logs and age-appropriate consent practices could be developed now, while others such as truth, fairness and bias may not have one specific answer even in the long term. These need to be addressed as we begin on a path to assess governance mechanisms guiding the implementation and development of AI.

The implications are:

- Competition to create more sophisticated models will push developments into new territories with unknown emergent skills & capabilities.
- There are many more models on their way and there is no putting the genie back in the bottle.
- Setting the AI alignment to values is selective and manual process done by the AI owner.
- Current models are not education specific, not designed or controlled by Australian education systems.
- Legislation to protect people's data could serve to impact subsequent model developments and entrench current leaders.

Long term rebalancing: the alignment problem

As mentioned above, bias can be both a feature of what is in the data the LLMs have learnt from as well as in the way that systems have been developed, and we no longer have transparency into this. However, biases in AI are real and come as a core element of any of

[16] https://www.anthropic.com/index/constitutional-ai-harmlessness-from-ai-feedback
[17] https://en.wikipedia.org/wiki/Alternative_facts
[18] https://www.edresearch.edu.au/using-evidence/standards-evidence

these networks, and it will take time to learn how to evolve tools to guide or control this, leading to a long-term value alignment problem.

‘Value alignment problems arise in scenarios where the specified objectives of an AI agent don't match the true underlying objective of its users. The problem has been widely argued to be one of the central safety problems in AI”[19]

These ethical, educational and wider LLM model governance issues are complex and will likely require forms of cooperation across jurisdictions, nationally and internationally. As new jurisdictional rules are developed (such as the EU AI Law or the US voluntary conformance) an assessment of their fit will need to be weighed against Australia's education needs and realities of influencing global markets.

The next section of this paper explores the principles and approaches different governmental and educational systems are reporting as their as a methods for defining the ‘what’ of their desired alignments. We will cover many different organisations' definitions of what we want from AI.

However, the implication above is that our ability to meet the principles outlined is another step further on. I.e., the ‘how' to respond to all the stated principles requires additional well considered steps.

In the final section of the paper we look briefly at the elements that could assist in progressing policy development and the related governance requirements to support these principles. We examine these through a lens of what is possible in the short term and what’s needed for medium term as well as looking at any rebalancing activities to begin addressing the ‘how’ of values alignment.

[19] http://lcfi.ac.uk/projects/completed-projects/value-alignment-problem/#

3. Organisations and activities internationally

More than 300 AI policy initiatives exist from over 60 countries.[20] Whilst education is mentioned in some of these, it is normally only as part of a larger national AI strategy. UNESCO has identified three main approaches to AI policy responses.[21]

1. Independent approaches (stand-alone policy or strategy)
2. Integrated approach (integrating AI elements into existing Education or ICT policies)
3. Thematic approach (focussing in one topic such as data security)

These are then broken down over different types of organisations actively supporting the education sector.

Supranational organisations

European Commission

Legal initiatives

To accelerate investments in AI, make AI programmes and strategies actionable, and to ensure policy alignment across borders, the European Commission has developed a Coordinated Plan on Artificial Intelligence[22]. The Commission has also proposed three, inter-related legal initiatives that will contribute to building trustworthy AI[23]:

1. a European legal framework for AI[24] to address fundamental rights and safety risks specific to the AI systems by focussing on the classification of tools and applications,
2. a civil liability framework[25] - adapting liability rules to the digital age and AI which covers topics like compensation for damages caused by unsafe products,
3. a revision of sectoral safety legislation (e.g., Machinery Regulation[26], General Product Safety Directive).[27]

The legal framework, otherwise known as the EU AI Act will probably not become effective before 2026 and while essential, crafting new laws for emerging technologies is difficult and often time-consuming.[28] The risk-based approach to legislation has been met with criticism and a demand for a more nuanced approach, especially in the education sector and

[20] https://oecd.ai/en/dashboards/overview
[21] https://unesdoc.unesco.org/ark:/48223/pf0000376709/PDF/376709eng.pdf.multi
[22] https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence
[23] https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence
[24] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
[25] https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12979-Product-Liability-Directive-Adapting-liability-rules-to-the-digital-age-circular-economy-and-global-value-chains_en
[26] https://ec.europa.eu/docsroom/documents/45508
[27] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0346
[28] https://oecd.ai/en/wonk/roadmap-ai-learning-campaign

specifically because of uniquely high number of micro-organisations (1-10 employees) within Europe.[29]

Unified standards

Standards play a vital role in the development of education environments and will be imperative in supporting compliance with government regulations and guidelines as well as facilitating necessary discussions around standardisation. As an example, the European Standardisation Organisations, CEN-CENELEC are developing harmonised European standards which could provide a legal presumption of conformity to AI providers.[30] Topics which will be addressed within the CEN-CENELEC formats for AI systems include risk management, governance, and quality of datasets used to build AI systems, record keeping, transparency and information provision, human oversight, accuracy specifications, robustness specifications, cybersecurity specifications, quality management system for providers, including post-market monitoring process, and conformity assessment. Also at the international standards level, ISO has specific standards subgroups looking into digital education related issues which may also include AI in the near future.

It is important to note that all of these standards are voluntary instruments.

Council of Europe

The Council of Europe has initiated the Steering Committee for Education Policy and Practice (CDPPE), adopted guidelines on Children's Data Protection in an Education setting[31] and it has produced a revised strategy on the rights of the child.[32] The Council believes that, unlike AI ethics frameworks, human rights are enforceable and, therefore, more fitting to govern AI throughout its life cycle.[33] The central idea behind a Human Rights based approach has already been introduced through the General Data Protection Regulation (GDPR) and is reflected in the draft of the Convention on Artificial Intelligence, Human Rights, Democracy and the Law.[34]

In 2022 the Council of Europe released a report on AI and Education through the lens of human rights, reviewing connections between AI, education, and the challenges to human

[29] https://static1.squarespace.com/static/5fac2fdb0da84a28cc76b714/t/63bfda44de4b365544ae4b45/1673517650701/EEA+Edtech+Map+Insights+Report+2022.pdf

[30] https://publications.jrc.ec.europa.eu/repository/handle/JRC132833

[31] Consultative Committee on the Convention for the protection of individuals with regard to automatic processing of personal data, Convention 108, Guidelines: Children's Data Protection in an Education setting, November 2020.

[32] https://www.coe.int/en/web/children/strategy-for-the-rights-of-the-child

[33] https://policyreview.info/articles/analysis/future-proofing-the-city referencing Donahoe & Metzker, 2019; McGregor et al., 2019; Yeung et al., 2020; Smuha, 2020; Cobbe et al., 2020

[34] Council of Europe. (2023). Committee on artificial intelligence: Revised zero draft [framework] convention on artificial intelligence, human rights, democracy and the rule of law (CAI(2023)01; pp. 1–13). https://rm.coe.int/cai-2023-01-revised-zero-draft-framework-convention-public/1680aa193f

rights.[35] The Council of Europe is currently preparing a recommendation to go to all member states for a legal instrument to specifically govern the development and use of AI in education.

UNICEF

UNICEF has developed a rubric entitled 'foundations for child centred AI'. These foundations encompass the support of children's development and well-being, ensuring inclusion, prioritising fairness and non-discrimination, protecting children's data and privacy, ensuring safety for children, providing transparency, explainability and accountability, empowering governments and businesses with knowledge of AI and children's rights, preparing children for present and future AI developments, and creating an enabling environment.[36]

Highlighting key risks and opportunities, UNICEF also looks at how to create inclusive AI[37] and create awareness for the fact that reaching the age of digital consent doesn't mean our youth should be digitally treated like adults.

While making general recommendations for the holistic creation of an appropriate ecosystem for fostering child-centred AI, UNICEF acknowledges that requirements will differ depending on local contexts without providing any further recommendations for regulatory mechanisms or methods of recourse or how to get to these steps. UNICEF does make support resources for parents and teens, and a road map for policy strategy development available.[38]

UNESCO

UNESCO has developed support resources at all levels of the education ecosystem publishing documents on Generative AI and the Future of Education,[39] Artificial Intelligence and the Futures of Learning,[40] and a mapping of government endorsed AI curricula.[41] UNESCO will soon be launching guidelines on the use of Generative AI in education (and research) through the work of an expert group, and a draft of frameworks of AI competencies for teachers and students (for publication in early 2024).

Intended to help policy makers in the development of their strategies, UNESCO makes several recommendations in their paper, AI and Education: Guidance for Policy Makers.[42] These include actions like developing a master plan, fostering local AI innovation, assessing system readiness, and choosing strategic priorities. However, there is no guidance provided regarding possible enforcement mechanisms or any international agreement or alignment.

---

[35] https://book.coe.int/en/education-policy/11333-artificial-intelligence-and-education-a-critical-view-through-the-lens-of-human-rights-democracy-and-the-rule-of-law.html

[36] https://www.unicef.org/globalinsight/reports/policy-guidance-ai-children

[37] https://www.unicef.org/globalinsight/stories/developing-girls-digital-and-ai-skills-more-inclusive-ai-all

[38] https://www.unicef.org/globalinsight/media/1166/file/UNICEF-Global-Insight-tools-to-operationalize-AI-policy-guidance-2020.pdf

[39] https://unesdoc.unesco.org/ark:/48223/pf0000385877)

[40] https://www.unesco.org/en/digital-education/ai-future-learning)

[41] https://unesdoc.unesco.org/ark:/48223/pf0000380602)

[42] https://unesdoc.unesco.org/ark:/48223/pf0000376709

OECD

The OECD has collated an overview of existing AI policy Instruments[43] globally spanning hundreds of documents. Based on the 2019 G20 AI Principles,[44] the OECD has developed five principles and five recommendations for policy makers[45] covering transparency, explainability, accountability, human in the loop, and protecting data. Believing that AI's biggest promise lies in the personalisation of learning and learning materials,[46] the OECD has identified that the biggest challenge is creating and maintaining trust.

To both increase the understanding about AI and support the skills acquisition necessary for new workforce needs, the OECD suggests a need for a global AI learning campaign.[47] This would cover both learning about and learning to learn and work with AI. Highlighting the need for collaboration to ensure the success of policy measures, they identify that multi-stakeholder collaboration should increase the likelihood that the resulting AI laws and policies will successfully protect individuals.[48]

National (governmental)

In a review of over 88 frameworks or guideline documents, it was found that more than half were created by the public sector and these were primarily engaged in differentiating themselves both from allies and opponents[49] making minimal reference to frameworks such as IEEE's Ethically Aligned Design[50] or the Toronto Declaration.[51] Globally, governments and ministries understand the need for quick action and the ramifications of inaction but are finding it challenging to proceed from principles and guidelines through to enforceable measures. There is, however, also an increasing effort to align or work together with other nations as, "[e]ven if resources related to AI are concentrated in a specific country, we must not have a society where unfair data collection and infringement of sovereignty are performed under that country's dominant position."[52]

National strategies on AI are being developed around the globe. These are comprehensive and locally applicable documents and whilst a small number of these explicitly include information about AI practices for education, this is not always the case. The United States'

---

43 https://oecd.ai/en/dashboards/overview/policy

44 https://www.oecd.org/education/trustworthy-artificial-intelligence-in-education.pdf

45 https://oecd.ai/en/ai-principles

46 https://www.oecd.org/education/trustworthy-artificial-intelligence-in-education.pdf

47 https://oecd.ai/en/wonk/roadmap-ai-learning-campaign

48 https://oecd.ai/en/wonk/roadmap-ai-learning-campaign

49 Daniel Schiff, Justin Biddle, Jason Borenstein, and Kelly Laas. 2020. What's Next for AI Ethics, Policy, and Governance? A Global Overview. In 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES'20), February 7–8, 2020, New York, NY, USA. ACM, New York, NY, USA, 6 pages. https://doi. org/10.1145/3375627.3375804

50 https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

51 https://www.torontodeclaration.org/

52 This is Principle 4.1.5. Principle of Fair Competition in Japanese Cabinet Office, Council for Science, Technology and Innovation, 'Social Principles of Human-Centric Artificial Intelligence' (2019) <https://www8.cao.go.jp/cstp/english/humancentricai.pdf>

Blueprint for an AI bill of rights[53] is not enacted policy but will likely lay the groundwork for important decisions and policy development. It identifies five principles and associated practices; safe and effective systems, algorithmic discrimination protections, data privacy, notice and explanation, human alternatives, consideration, and fallback. Its progress is complex and has been complicated by the work of many other departmental based AI principles and by lack of agreement on scope.[54]

The Office of Education Technology (USA) offers significant resources on AI[55] and its recent report on AI and the Future of Teaching and Learning[56] and Core Messages[57] offer useable insights and outline best practice recommendations, including emphasising humans-in-the-loop, prioritising trust, focussing on R&D for context and safety, and involving and informing educators.

Other, fast-moving government initiatives include China's preparation of the next-generation AI-workforce through the implementation of AI education programmes from K-12 through to post-secondary level education and relying on public-private partnerships58. To this end, the Next Generation Artificial Intelligence Development Plan was developed with the goal of making China the world's primary innovation centre by 203059. Similarly, being one of the globally fastest-growing economies, India's national strategy for Artificial Intelligence shows its plan to become a global AI leader and addresses the key barriers that need to be addressed in order to achieve this[60.]

Government endorsement of voluntary commitments

The Biden administration has worked with US industry to create a set of voluntary commitments from leading Artificial Intelligence Companies to manage risks. [61] Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI agreed on July 21st to work together on three areas, Safety, Security and Trust.[62]

1. **Ensuring products are safe before introducing them to the public**
   This involves security testing and the sharing of information to mitigate risks
2. **Building systems that put security first**
   Resulting in an investment in cybersecurity and third-party reporting

---

[53] https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

[54] https://www.brookings.edu/articles/opportunities-and-blind-spots-in-the-white-houses-blueprint-for-an-ai-bill-of-rights/

[55] https://tech.ed.gov/ai/

[56] https://tech.ed.gov/ai-future-of-teaching-and-learning

[57] https://tech.ed.gov/files/2023/05/ai-report-core-messaging-handout.pdf

[58] **Nurturing the Next-Generation AI Workforce: A Snapshot of AI Education in China's Public Education System**, Published:March 7, 2022, Author: Xiaoting (Maya) Liu: Project Manager, Risk Analysis & Development

[59] Jia He, "The Next Generation AI Development Plan — What's Inside?," medium.com, August 10, 2017, https://medium.com/@jiahe/the-next-generation-ai-development-plan-whats-inside-72824a9bcc3.

[60] https://niti.gov.inwritereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf

[61] https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/

[62] https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf

3. **Earning the public's trust**
   Including developing technical mechanisms ensuring users know when they are interacting with an AI, a commitment to publicly report on AI systems' capabilities, limitations, and appropriate and inappropriate use, the prioritization of research on risks, and a focus on using the technologies to help address society's greatest challenges.

These three principles provide guidance regarding key issues and topics that urgently need to be addressed and which are technically able to be addressed. This concept aligns with the idea of 'base safety' outlined in the final section of the paper and we urge a similar practical approach.

This is a landmark accomplishment, as it means the US vendors have been involved in the definition of and aggreging to some benchmarks, however these are a limited set and do not address most of the ethical principles outlined in the previous section. Like many others, it's also a voluntary code of practice and without accountabilities or financial penalties.

However importantly it is achievable in the near term, and it gives the intentions gravitas. It has set a direction in motion. A question however is to what degree is 'voluntary self-regulation' enough? Later we discuss the risk-based approach from the EU that is, comprehensive but burdensome and provides recourse and paths to penalties. However it suffers from being built outside the industry and if the US approach works it could undermine the viability of the EU approach.

Local/regional (agencies)

A number of countries work with an agency support model similar to that of ESA in that an agency is directly aligned with the ministry or ministries and is tasked with supporting the analysis of new concepts and ideas as well as the development and implementation of practical tools to accompany the digitisation of the education systems in their region. Some examples of these are outlined below.

Digital Agency Brandenburg, Germany

An integral part of the state's strategy for digital transformation, the Digital Agency Brandenburg[63] (DABB) is funded by the state ministry for education and the state investment bank. Their activities range from initiating and implementing digital projects with high state importance, holistic support of the municipalities in their strategic planning and operative implementation of technologies, and leading digitisation projects.

The DABB and the state of Brandenburg as leaders of the national taskforce on AI in Education are in the process of evaluating an AI based intelligent tutorial system, which could be implemented nationally, as well as a strategy using AI to cover gaps appearing due to the severe and increasing shortage of teachers, specifically in more rural or remote areas.

63 https://www.digital-agentur.de/digitalagentur#c85

They are supporting the combined state ministries in the development of an AI policy document.

IKT Norge, Norway

ICT-Norway creates both general framework conditions for the entire industry and identifies specific questions that could be difficult or challenging for an individual company to address alone. Their approach with AI ranges from giving all political offices within the country a mandatory summer reading list including the book “machines that think”, to leading the latest public enquiry regarding the use of AI in schools. 130 million Krone (approximately 20 million AUD) have been set aside as a budget for the innovation and development of AI practices and tools for education.

Educa, Switzerland

Educa is a specialist agency commissioned by the Confederation and cantons. It combines technological developments with quality improvements in the Swiss education area. They explore the ecosystem, mediate between the various stakeholders and establish new services. Their current focus is on further developing sound data practices and regulations which will also affect the use of AI in education (e.g. data usage policies, data governance strategies, and a unified digital identity approach[64]).

Jet Educational Services

JET works with government and the public sector, civil society organisations, local and international development agencies and educational institutions to improve the quality of education and the overall relationship between education, skills development and the world of work within South Africa. JET’s work with AI stems from their strong work developing interoperable systems and interoperability standards and data frameworks. Building on their review of the pan commonwealth standards framework for teachers and school leaders[65], they have a strong focus on developing training resources and possibilities for teachers and their skills development[66], with emphasis on AI for learning, project based and next generational skills.

Departmental or initiative based principles

Some countries (like the UK and Australia) have different parts of government creating similar but not identical principles for different sectors, each purporting to encompass all aspects of AI as well as those specific to the sector. A good example in Australia is found across;

- Australia’s Artificial Intelligence Ethics Framework from Industry dept,
- The Mandatory Ethical Principles for the use of AI from NSW’s Digital
- Schools AI framework for Education

64 https://www.educa.ch/en

65 https://www.jet.org.za/clearinghouse/projects/primted/standards/general-standards/pan-commonwealth-standards-framework-for-teachers-and-school-leaders.pdf/view

66 https://www.jet.org.za/resources/understanding_the_impact_of_ai_on_skills_development.pdf/view

Each framework is similar but not the same and for them to be implemented will require some enforceable rules or voluntary commitments. Furthermore, they are likely to apply to many of the same industry players who are also more likely to be overseas based.

NGOs (civil sector)

The civil sector has been quick to respond to the struggle that schools and educators are facing to keep up with not only the pace of change, but also an understanding of what is required to ensure safe and equitable education environments. This can be seen in the resources currently being developed.

ISTE

The International Society for Technology in Education (ISTE[67]) has created targeted resources to support school leaders better understand AI in educational settings and how it can be practically implemented[68]. For teachers it has developed 15-hour Professional development course to help teachers explore AI[69], and hands on guides to help engage students across elementary, secondary and in elective's computer science and ethics. And podcasts featuring the real-world experiences of many schools.

CoSN

The Consortium for School Networking (CoSN[70]) in their response to Artificial Intelligence and Generative AI[71], focus on essential leadership guidelines, training requirements, developing school policy and integrating or leveraging existing privacy and security measures. CoSN's report "Artificial Intelligence in K-12"[72], presents a model for including learning inside a wider map of AI impact on society. They include the centre for curriculum reforms 4-dimensions of 21st century skills, reinforcing a trend that links how AI is increasing the need for education to recognise and adopt these new skills. It also raises the need for data protection to be actioned now.

Digital Promise

Having supported the Office of Educational Technology in the USA develop their insights and recommendations document, Artificial Intelligence and the Future of Teaching and Learning[73], Digital Promise takes a leadership role as a major R&D institute, supporting policy creation for the US government, hosting product certifications that serve industry, and

[67] https://www.iste.org/about/about-iste

[68] https://cdn.iste.org/www-root/2023-07/Bringing_AI_to_School-2023_07.pdf?_ga=2.152508027.319459478.1689974430-11529736.1689974430

[69] https://iste.org/professional-development/iste-u/artificial-intelligence

[70] www.cosn.org

[71] https://www.cosn.org/wp-content/uploads/2023/04/EmpSupAI.pdf

[72] https://www.cosn.org/tools-and-resources/resource/artificial-intelligence-ai-in-k-12/

[73] https://tech.ed.gov/files/2023/05/ai-future-of-teaching-and-learning-report.pdf

supporting educational leaders and practitioners. Digital Promise has access to some leading AI education experts, evidenced by their head of AI having developed the US office of ed tech's AI report and authors of OECD reports writing their blogs. There landing pages[74] blogs[75] and partnerships such as engage AI[76] create accessible summaries on topics like AI literacy for educators, automation with Gen AI.

Institute of Analytics

Directly hoping to support policy makers and governments, the Institute of Analytics, a non-for-profit dedicated to harnessing the power of data analytics, has developed a checklist for an organisational generative AI policy[77] as well as Model cards[78] to help with the categorisation and development of AI tools.

Center for Democracy and Technology

Likewise, the Center for Democracy and Technology provides information and overview of general developments in AI[79] to support knowledge exchange. The initiative AI4K-12[80] seeks to develop national guidelines for AI education in K-12, which are aligned around 'Five big ideas' in AI[81], which is a similar model to a principle approach.

Teacher and student focused initiatives

A comparative study of AI curricula globally found that, to date, there are only eleven AI curricula which have been developed and implemented by the governments[82] of Armenia, Austria, Belgium, China, India, Republic of Korea, Kuwait, Portugal, Qatar, Serbia and UAE. This is sobering considering the world's citizens need to understand what the impact of AI might be, what AI can do and what it cannot do, when AI is useful and when its use should be questioned, and how AI might be steered for the public good.[83]

Whilst there are a number of grassroots initiatives with teachers sharing lesson ideas, overviews of best tools, or even creating entire free courses[84] to help each other, there are also a number of more systematically designed resources becoming available for teachers and students.

---

[74] https://digitalpromise.org/initiative/artificial-intelligence-in-education/

[75] https://digitalpromise.org/our-blog/?p=1&initiatives=&topics=artificial-intelligence

[76] https://engageai.org/

[77] https://ioaglobal.org/pdf/Checklist-IoA-Generative-AI-Policy.pdf

[78] https://ioaglobal.org/insights/model-cards-to-support-responsible-ai/#

[79] https://cdt.org/insights/?keyword=AI&area-of-focus%5B%5D=ai-machine-learning#results

[80] https://ai4k12.org/

[81] https://ai4k12.org/wp-content/uploads/2022/01/AI4K12_Five_Big_Ideas_Poster_3_19_2021.pdf

[82] https://unesdoc.unesco.org/ark:/48223/pf0000380602

[83] Miao, F., Holmes, W., Huang, R. and Zhang, H. 2021. AI and Education Guidance for Policy-makers. Paris, UNESCO. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000376709.

[84] https://kiik.ch/online-kurse.html

Civil sector response

The AI curriculum study received a total of 31 responses from NGOs stating that they had developed an AI curriculum for education.[85] With high profile financial backing, AIEdu[86] is a non-profit that provides a curriculum which creates equitable learning experiences and builds a foundational AI literacy. Their resources are aimed to support both teachers and students providing professional development and learning toolkits as well as a number of articulated projects.

ISTE has created the Hands On AI projects, which enable student-driven projects in developing AI and include multilingual guides.[87] With a focus on teachers and teacher training, ISTE also provides guides for teachers[88] and courses on understanding AI[89] and exploring the practical implementation of AI in the classroom.[90] Additionally, ISTE will soon be launching Stretch AI, a chatbot purely designed for educators to get a better understanding of ISTE Standards and the ASCD's research based pedagogical practices.[91] ISTE is also a founding member of the newly minted multi stakeholder 'Teach AI',[92] which is being run by Code.org and looks to create practical support mechanisms to bring policy into implementation through courses, infographics, workshops and viral information pieces. Additionally, they aim to revise the existing computer science frameworks to include AI practices and make these understandable for teachers and students. They are currently defining their strategy. From AI learning series, a glossary of key AI terms to interactive listening sessions, Digital Promise is working to inform and empower educators, administrators and decision makers within education.[93] Getting Smart is an educational technology publication which has created a significant resource walking teachers through more general AI concepts.[94]

The Association for the Advancement of Artificial Intelligence[95] offers education focused workshops and also sponsors AI4K-12, which aims to develop an online, curated resource Directory to facilitate AI instruction, and a community of practitioners, researchers, resource and tool developers focused on the AI for K-12 audience. AI4Ed promotes workshops, conferences and information sessions on AI in education.[96] They created on their five big ideas to structure the curriculum around teaching AI in schools.

---

[85] ttps://unesdoc.unesco.org/ark:/48223/pf0000380602_page 20

[86] https://www.aiedu.org/

[87] https://www.iste.org/areas-of-focus/AI-in-education

88 https://cdn.iste.org/www-root/2023-07/Bringing_AI_to_School-2023_07.pdf?_ga=2.259212169.264969171.1690704532-11529736.1689974430

[89] https://www.iste.org/learn/iste-u/artificial-intelligence

[90] https://www.iste.org/professional-development/iste-u/artificial-intelligence?_ga=2.207141399.543074122.1679943175-1137964326.1679449353

[91] https://info.iste.org/stretch-ai

[92] https://teachai.org

[93] https://digitalpromise.org/2023/06/06/supporting-ai-literacy-for-educators-new-and-emerging-resources/

[94] https://www.gettingsmart.com/whitepaper/artificial-intelligence/how-did-we-get-here/

[95] https://aaai.org/

[96] https://ai4ed.cc/

Ai412.org 5 big ideas

Resources and curriculum for teaching AI

AI Fluency	AI bootcamp
Ready AI	Kits & competitions
Invent XYZ	Real world projects & environments
AI4All	Education and mentorship
AI+Ethics	Middle grades course from MIT media lab (~30 hours)
AI for Oceans	Tutorial from code.org
AI Experiments	Experiments with Google
Teachable Machines	An initiative from Google

Public sector responses

Resources are currently being developed for teachers as part of their training or education. These follow two distinct directions: Learning about AI and learning to teach with AI. There is an urgent move to address the gap between the skills that have been identified as necessary and the curriculum resources being made available. OECD has recognised that students will not only need to be able to use AI tools, but also need diverse skill sets enabling them to be flexible and adapt to technological changes[97].

As a consequence, school networks and districts (USA) are now determined to support teachers and students embrace the potential of AI, such as the New York City Public Schools, which, despite being extremely cautious initially, are now creating a repository of learnings and findings and a community space to share these.[98] Additionally, NYC Public Schools will be expanding their ongoing Computer Science for All[99] initiative to encompass AI related resources.

Also directed at teachers and students, the Day of AI curriculum,[100] a project of MIT, helps teachers and educators to run activities within their classrooms and provides entire curriculum and training packages to facilitate this.

Private sector

[97] OECD (2023), *Is Education Losing the Race with Technology?: AI's Progress in Maths and Reading*, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/73105f99-en.

[98] https://ny.chalkbeat.org/2023/5/18/23727942/chatgpt-nyc-schools-david-banks?oref=csny_firstread_nl

[99] https://infohub.nyced.org/in-our-schools/programs/computer-science-for-all-overview

[100] https://www.dayofai.org/curriculum

Following on from successful programmes such as “Intel Teach”, which reached over 15 million teachers in 70 countries,[101] Intel has produced their Global AI Readiness Program,[102] Microsoft has developed a Learn AI Skills Challenge[103] and IBM the EdTech Youth Challenge.[104] Globally, smaller companies are also starting to develop resources for teachers, such as the German Fobizz[105] and for students, the Indian initiative Coding and More,[106] Dell and AWS are also entering the Gen AI and education sector with intentions to create resources.

Additionally, tech organisations have developed principles, such as the Microsoft Responsible AI principles[107] or IBM Ethical AI Principles.[108] These generally reflect the same set of principles as the other AI and ethical principle sets. Some of these, however, such as Google AI principles[109] also include distinct developmental guidance such as the need to be socially beneficial, upholding standards of scientific excellence and a do no harm directive.

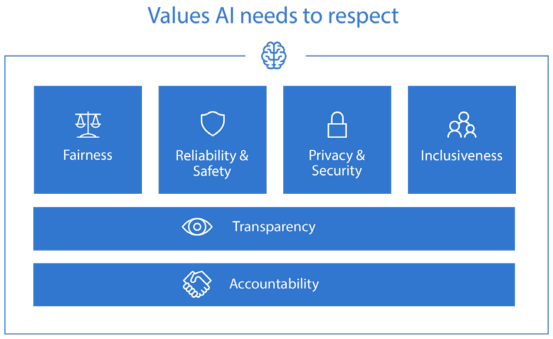
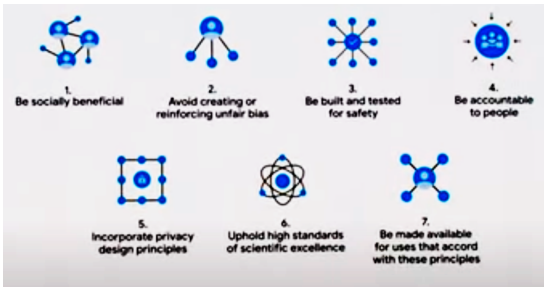


Chart 5.
Source: Microsoft Corporation



Googles 7 principles of Responsible AI

Industry-led

In April 2023, hundreds of Industry initiatives and researchers called for a pause in the development of AI technologies as “Powerful AI systems should be developed only once we are confident that their effects will be positive and their risks will be manageable.”[110] The main demand is that this pause can be used for the development of much needed policy to ensure the safety of these systems. These policy recommendations include the need to:

1. Mandate robust third-party auditing and certification.
2. Regulate access to computational power.

101 https://www.intel.com/content/www/us/en/homepage.html
102 https://www.intel.com/content/www/us/en/corporate/artificial-intelligence/digital-readiness-home.html
103 https://www.microsoft.com/en-US/cloudskillschallenge/ai/registration/2023?ocid=aisc23_CSC_skillsforaiblog_cnl
104 https://www.ibm.com/impact
105 https://fobizz.com/
106 https://codingnmore.com/
107 https://www.microsoft.com/en-us/ai/our-approach
108 https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf
109 https://ai.google/static/documents/ai-principles-2022-progress-update.pdf
110 https://futureoflife.org/open-letter/pause-giant-ai-experiments/

3. Establish capable AI agencies at the national level
4. Establish liability for AI-caused harms.
5. Introduce measures to prevent and track AI model leaks.
6. Expand technical AI safety research funding.
7. Develop standards for identifying and managing AI-generated content and recommendations.

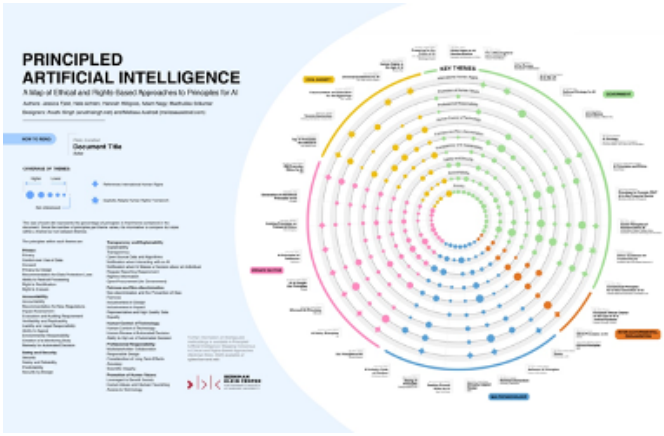
The issue of note here is that there is a manifest difference between the focus of these industry led initiatives and the principles identified across the globe from policy makers. The key difference is that this industry led framework preference near-term actionable steps over long term desired outcomes or intentions. That is focussing on the how over the what.

Ethical frameworks

Principles are being used to guide strategy and policy, but their implementation or adoption is voluntary and non-binding. Although hundreds of different principle documents have been created, they all share similar if not identical categories.

Best practices in mapping Ethical and Rights-based Approaches to Principles for AI[111] has been conducted by the Berkman Klein Center, which identified eight core categories of such principles:

1) Privacy
2) Accountability
3) Safety and Security
4) Transparency and Explainability
5) Fairness and Non-discrimination
6) Human Control of Technology
7) Professional Responsibility
8) Promotion of Human Values



These key themes have persisted without much change into further recommendations and guidelines. They were formed from an assessment of over twenty sets of principles from industry and government and were expertly segmented such that its categories have a minimum of overlap and a higher degree of alignment. This ensures that those who are to enact or be responsible for the enactment of a principle will have consistency within each category. This structure is what we recommend using when mapping the principles that have been developed across other organisations. It has served as a bases when seeking to find

[111] Fjeld, Jessica and Achten, Nele and Hilligoss, Hannah and Nagy, Adam and Srikumar, Madhulika, Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI (January 15, 2020). Berkman Klein Center Research Publication No. 2020-1, Available at SSRN: https://ssrn.com/abstract=3518482 or http://dx.doi.org/10.2139/ssrn.3518482

alignment. The Appendix contains a more detailed breakdown of this framework plus a selection of other principles frameworks.

The OECD AI principles[112] provide a number of examples to aid local understanding and implementation. They only go as far, however, as stating that AI systems should include a "values alignment" but do not look further into this. Additionally, these principles are still generic due to their encompassing definitions of AI and lack of segmental focus on education. With a clear difference in focus, the Montreal Declaration[113] raises the issues of AI technologies affecting life, the quality of life and the reputation of people and suggest certain practices to support this.

Best practice in education frameworks

Education specific frameworks like that of the Russel Group Universities in the UK[114] cover topics of AI literacy, use and impact, however, they scope these within the bounds of what staff within the University can impact. For instance "Universities will adapt teaching and assessment" in response to supporting academic integrity are set within the bounds of the academic rigour for each course. The responsibilities and mechanisms referred to are within the scope of the staff's current achievable responsibilities. Further aspirational, moral or ethical judgments, or AI systems performance are not included as principles for their staff to be concerned with.

Ethical framework for education

The Ethical Framework for AI in Education,[115] created in 2020 in partnership with Rose Lucken, the Nord Anglia school network and funded by Pearson, Microsoft and McGraw Hill sets out nine ethical objectives and, and develops criteria for these. The U.S Department of Defence adopted Ethical AI principles expanding the usual list of principles with the idea of needing Governable AI.[116]

Ethical principles are difficult to legislate

The strong focus on ethics has stemmed from public concern and a current lack of evidence regarding the societal effects of AI. Comparative studies suggest, though, that ethics guidelines and frameworks can be used to convince legislators that stakeholders can self-govern and that specific legal instruments are not necessary[117]. An analysis of 22 major ethics guidelines even highlights that "AI ethics—or ethics in general—lacks mechanisms to reinforce its own normative claims" and that principle frameworks like this "are rather weak

[112] https://oecd.ai/en/ai-principles

[113] https://declarationmontreal-iaresponsable.com/wp-content/uploads/2023/01/UdeM_Decl-IA-Resp_LA-Declaration-FR_vFINALE_2_j.pdf

[114] https://russellgroup.ac.uk/news/new-principles-on-use-of-ai-in-education/

[115] https://www.buckingham.ac.uk/wp-content/uploads/2021/03/The-Institute-for-Ethical-AI-in-Education-The-Ethical-Framework-for-AI-in-Education.pdf

[116] https://www.defense.gov/News/Releases/release/article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/

[117] Calo, R. (2017). Artificial intelligence policy: a primer and roadmap. SSRN Journal, 1–28.

and pose no eminent threat"[118] to any of the stakeholders thus not bringing about the change that they suggest is necessary.

Research of ethical frameworks

Researchers continue to create summaries of different principles frameworks. A good summary comparison of various principles from non-governmental organisations is Ethical Principles for Artificial Intelligences in K-12 Education.[119] There have been many studies worthy of noting, however the time lag means their results often do not include Generative AI considerations. Two very worthy meta-analytics to review, however, include Ethical principles for artificial intelligence in K-12 education[120] & Generative AI: Implications and Applications for Education.[121]

Summary table

Type of resource	Example Organisations	Example Resources	Type
Legal Instruments	EU Commission Council of Europe	EU AI regulatory Framework, AI Act, Convention on Artificial Intelligence, Human Rights	Binding formal
International Frameworks	IEEE,	Ethically aligned design, Toronto Declaration, Montreal Declaration	Semi-Binding
National Strategies	USA	Blueprint for an AI Bill of Rights	Binding
Industry Voluntary Framework	USA	"Ensuring Safe, Secure, and Trustworthy AI"	Voluntary
AI Guidelines	UNICEF	foundations for child centred AI	Voluntary
AI Policy Guidelines	UNICEF, Institute of Analytics	Tools to operationalise AI policy Guidance, Checklist for an organisational generative AI policy, road map for policy strategy development	Voluntary
AIEd Frameworks	AI4K-12, Office of Education Technology (USA)	National Guidelines for AI in Education, Artificial Intelligence and the Future of Teaching and Learning	Voluntary
Standards	CEN-CENELEC, ISO		Voluntary
Learning Recommendations	OECD	Global Learning Campaign,	
School leadership resources	ISTE, COSN, Digital Promise, AI4Ed, NYC Public Schools	Bringing AI to Schools, Artificial Intelligence and Generative AI, Repository and community	Voluntary
Teacher resources	ISTE, COSN, Other teachers, AIEdu, Teach AI, Digital	Hands on Projects, AI Edu Curriculum, Understanding AI teacher guide, Practical	Voluntary

[118] Hagendorff, T. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds & Machines* **30**, 99–120 (2020). https://doi.org/10.1007/s11023-020-09517-8

[119] https://doi.org/10.1016/j.caeai.2023.100131

[120] https://www.sciencedirect.com/science/article/pii/S2666920X23000103

[121] https://arxiv.org/abs/2305.07605

	Promise, AAAI, NYC Public Schools, MIT, Intel,	implementation guides, Chatbots for teachers, Teach AI, Day of AI curriculum, Digital Readiness Programme	
Student resources	ISTE, AIEdu, Teach AI, MIT, Intel, IBM	Hands on Projects, AI Edu Curriculum, Day of AI curriculum, AI for Youth, Youth Challenge	Voluntary
Parent resources	UNICEF	Support resources for parents	Voluntary

Principles section summary

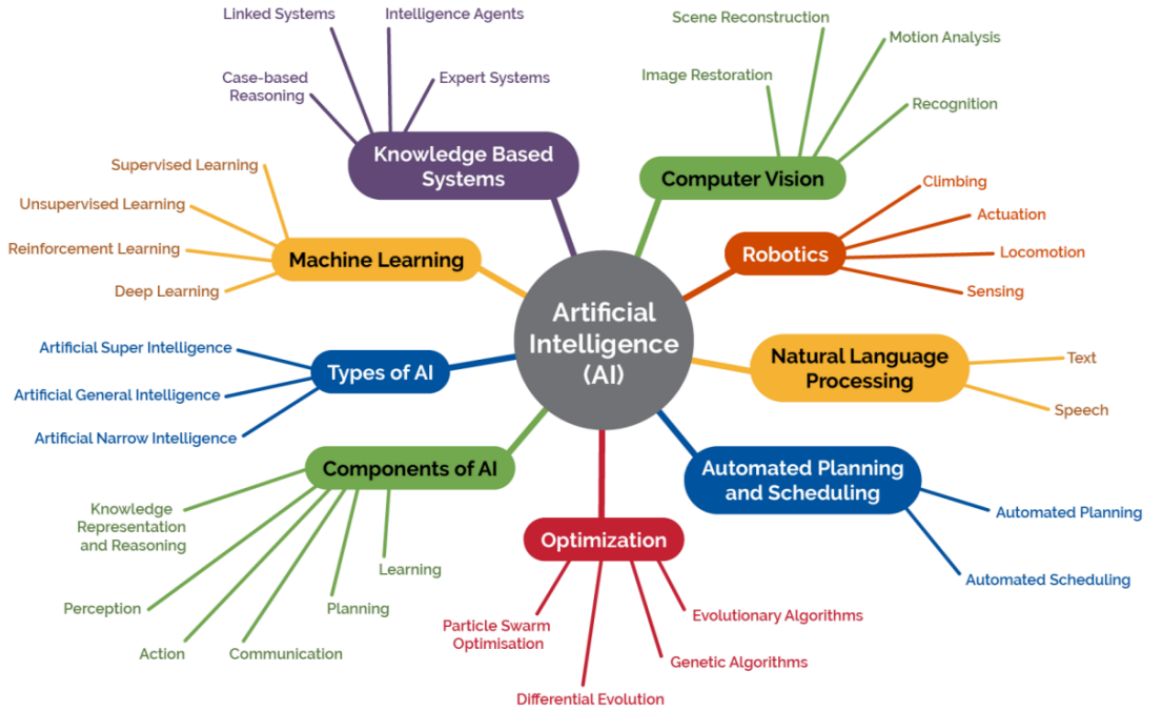
Hundreds of principles for the implementation and development of AI have been created by many stakeholders across all levels of the education landscape. These principles predominantly cover the same issues from transparency to explainability, safety to no bias that the Berkman framework laid out two years ago. Few are on track to be or can currently be legislated. This stems from the fact that these frameworks are lacking mechanisms to enforce them. The success of either the EU's legal path or the U.S. voluntary path will likely set the overall direction for national ethical AI enforcements, however, this will take some time to play out.

Very few of these principles specifically cover any of the unique issues related to child safety, child and youth development or the education sector's requirements. Although principles explicitly created for the education sector do expand this list, they generally do this only by adding principles including "aligned to vision for learning" that are again laudable, but with limited scope for legislative action. National AI policy strategies with a focus on Education **for** creating AI or education for society to understand AI, are typically general and do not focus on AI for Education or education specific needs of AI.

At times, there appears to be an assumption that stating principles will lead to a form of governance or self-governance. Industry based principles that exist are in alignment with a subset of the government-initiated principles. From the authors' interviews, though, there does not appear to be a strong movement to adopt any one department or country's principles. There does, however, appear to be an explanation to policymakers that the current industry alignment is sufficient such that specific legal instruments which direct financial accountability are not necessary.

The majority of policy guidelines or other resources developed for supporting AIEd strategy follow a similar approach to the principles papers, highlighting a good high-level or general framework, but not providing information on how to take issues from being defined as principles to being implemented and enforced in the local learning environments. This high-level approach, whilst listing important goals for educational spaces and implementations of AI, often misses some of the key issues that require regulation and enforcement the most and remains difficult to lead through to some form of implementation.

4. Current AI ecosystem



Office of Educational Technology original AI map.

Diagrams that describe the traditional technical landscape of AI like this are useful, however they quickly become out of date as new LLMs evolve into multimodal (that is, many data types) foundational models. In this section we look at how the market ecosystem is evolving, how the until recently separate segments of AI research, like above are combining in the LFM segment, how open source and government investments are splitting off to create smaller use case specific versions and how both the education enterprise market and the vendors' applications are also incorporating LLM capabilities.

The Machine Learning from previous AI's is being folded into LFMs. The traditional development of an AI model works by a machine learning algorithm's overlay put on top of a sufficient data set to evolve out a specific probability network which automates responses. That is, traditionally we used raw training data, create a model and then use that model for say recommending music or movies, or to recognise a face in a picture.

That changed with the breakthrough in the "attention is all you need paper"[122] providing the bases to build a conceptual foundation and now data sets are added by seeking commonalities between the already learning concepts and new data. For new data sets such as pictures we call this Multimodal: the LLM can understand the concept represented in a picture as well as in a sentence written as a prompt. Its strength is that it overlaps the concepts expressed in a picture and those in the paragraph and so the massive learning already done is added to by these new data sources. And it turns out that this extends to

[122] https://arxiv.org/abs/1706.03762

concepts in sensing, movement as well as written words, visual comprehension and auditory comprehension.

Large foundational models industry

The USA leads in the development of foundational technical models. There are about 12 organisations responsible for the majority of the Gen AI with new entrances such as Apple's AJAX[123] and X.AI[124] from Elon Musk. However, there are over 100 versions of these LLMs.

It is likely that there will be more for-profit organisations emerging, however, as time goes on there will be diminishing returns to create models at the scale that can compete as these foundational models move toward higher levels of capability. What this means is there will be less market for new entries and it will be more difficult to compete so there will be a finite set of the largest and most capable models.

LFMs are not universally applicable

Multimodal Large Foundational Models are a massively expensive and complex thing to create. As such it's not right for all circumstances which can lead to a new segmentation of the market where smaller and open-source models are being used to create intelligent situation specific models. The LLM/LFM AI ecosystem is now bifurcating along market lines, with smaller models being 'fine-tuned' to a specific need.

This is done by taking an open LLM model (the most famous is Llama from Meta) and changing it just enough to be specific for a use case. There are now many thousands of these experimental fine-turned versions of open source LLMs on a service called huggingface[125] which operates a leader-board for the best models. They claim over 250,000 AI models and 100,000 applications using them.

To use these fine-tuned models they need to be set up and hosted, so are expensive and not something schools will do, however vendors are such as Pearson and MerlynMind, have already created their own versions of LLMs.

We are familiar with a chat bot (such as Chat GPT) which uses the GPT 3.5 (free) or GTP-4 core model. This is the consumer model for accessing the core model, however, new approaches are appearing for enterprise and for the online tech ecosystem.

[123] https://www.bloomberg.com/news/articles/2023-07-19/apple-preps-ajax-generative-ai-apple-gpt-to-rival-openai-and-google?in_source=embedded-checkout-banner

[124] https://www.newscientist.com/article/2382426-what-is-xai-elon-musks-new-ai-company-and-will-it-succeed/

[125] https://huggingface.co/

Consumer model: Direct to end user

Consumer Tool	Hosting	Model
Chat GPT	USA hosted on OpenAI	GPT- 3, 3.5, 4
Bing Search	USA hosted on Azure	GPT 3, 4
Google Search	USA hosted on Google	Palm1, 2
Google assistant	USA hosted on Google	LaMDA, Gemini

Application ecosystem

Application vendors are both reusing these foundational models and creating their own which can be from an unknown LLM or from a modification to a known LLM. The reason this matters is because with many applications using many LLM's the governance considerations around applying the principles becomes much more complicated.

Kahmingo	Tutor which helps student find and lean the answer rather than just give then the answer	GPT-4 highly modified
Bromcom	Ask your student management system what you need and it gives you help to achieve this.	Unknown
Notion	Create resources for meetings, project management	GPT-4, lightly modified
Pearson	Textbook tutor, knows all Pearson texts and helps students while they read.	Unknown
Talid	A teacher prompt enhancer with teacher PD assistant to lower barriers to use in schools	Google edu-tuned
MerlynMind	Whiteboard AI addon which allows teachers to ask the whiteboard to do common class tasks,	Open Source highly modified[126]

[126] https://huggingface.co/MerlynMind/merlyn-education-corpus-qa

Enterprise education

These are for the large education systems to turn on for their own applications. There are large education systems across the world examining options for LLM's for their use, both within the USA, and Australia.

They are mostly looking at taking one of the main LFM's hosted in the USA and adding 'context' to help make these more appropriate for education. The underlying models stays the same but it's given the extra context to make it more localised. This is being developed for all the major platforms. There are hosting options, e.g. Amazon to take the open source models and use those on their hosting. Google is trailing an enterprise layer where it can select from a range of the large providers (enabling selection of best practice LLM). Data sent and used is private and hosted in Australia and the enterprise can use their security systems and Microsoft is developing ways to use their models in education.

Government initiatives

The UK has announced an ambition to be competitive in Gen AI with significant investment in supercomputing capability (~1B GBP). Many counties are likely to desire a level of AI independence and local governance so we can expect a growth in state sponsored projects. Within countries, individual sectors (like education) or institutions (like universities) are considering model implementation options to support local management, data protection and governance. Further work is required to determine options for local governments to exert controls over US based systems, or the option to locally operate an Australian with an education specific model.

The long tail of Open Source and hosted LLM's

There are many thousands of smaller Open Models, which are being trained for niche specific tasks. These could be more efficient and cost effective for education vendors when they need this specialist task, rather than the larger models. We need to be aware that, in a market where the most effective approach is aligned with meeting customer needs, the inclusion of many smaller AI's could occur and the ability to know that these exist or their extent of use will be complex and jurisdictionally challenging.

Monitoring of many systems using many LLM's

The edtech monitoring system Learn Platform Edtech Report[127] estimates there are 9,000 tech tools in schools use across US, with each teacher using 42 different applications and large districts using a combined total of over 2000 separate tools.

[127] https://www.instructure.com/resources/research-reports/edtech-top-40-look-k-12-edtech-engagement-during-2022-23-school-year

As the proliferation of different models and the complexity of their interconnection increases, we need to consider what this means for the principles and frameworks which are to be applied. There are many burdens and risks to be mitigated when looking at safely using LLMs in education, the type of hosting and the source of the model alone could cause a level of complexity that would be difficult for schools to handle. There is a need to navigate these, making sure the work required by our education systems enacting principles also understands the nature of the market it is looking to have them implemented in. There are multiple players with a complex set of needs in a circumstance where a significant number of teachers and schools are already fully operationally using consumer-based approaches, that anecdotally is saving them time.

Existing monitoring system

Safer Technology for Schools[128] (ST4S) is an example of a government lead program which sets out clear technical expectations for industry and is linked to procurement hurdles to appropriately provide the framework for education specific companies to meet. The rules can be aligned to some of the existing global security standards, where clear measures and definitions for these exist. It has demonstrated an approach which has extended to many hundreds of Australian and international applications.

Summary

- There are many models, all of which are evolving at this time.
- There are known large models and many open-source or find-turned models, with thousands available,
- Vendors will use an array of different models to solve their needs and there are many vendors supporting schools
- With each model's new version release, the tests and conformance measures will have to be retaken, therefore how enactment or principles is implemented needs to factor in cost and complexity, with careful consideration on impacts on each part of the ecosystem
- There are enterprise models which are moving toward supporting some of the more clear and technically achievable principles.
- Some Governments are investing in creating and understanding the options available with this new ecosystem.
- There are some existing models to examine for their variability to undertake part of the enactment and compliance required.

[128] https://st4s.edu.au/

5. Getting from principles to policies and practices

Introduction

The above information and analysis show that there is substantial work going on with frameworks and principles and that they overlap in scope and ambition. There are also activities in the industry ecosystem that are occurring which could, in some cases, lead to the proliferation and complexity of the assumptions underpinning the principles and, in some cases, even challenge them. This means that all jurisdictions need to review how they translate these principles into manageable policies, guidelines, procurement advice, data protection updates, and guides for how systems go about implementing LLMs and LFMs in line with these. This review should also identify which principles schools themselves should develop policies for, which ones their education system will enact and support schools in and which are national, or the responsibility of the vendors providing the LLM.

Core challenges in getting principles into practices

The core challenges are:

- **Clarity** - of definitions, scope, jurisdiction etc.,
- **Measurability** - of what compliance with a principle is,
- **Enforceability** – the ability to act or to even know if compliance is occurring,
- **Urgency** - some principles need to be addressed more urgently than others.

Definitions of principles are not universal and there is overlap and difficulties, for example

- Different governments across the world have their own principles but the LLMs are global in access and use.
- Within a national government, different departments are developing their own principles. For example, what exists today in Australia is:
    - Australia's Artificial Intelligence Ethics Framework from Industry dept[129],
    - The Mandatory Ethical Principles for the use of AI from NSW's Digital[130]
    - Schools AI framework for Education[131]
- Is there an expectation that vendors will be required to adhere to each of these, or could it be argued that they have made a best effort by putting forward their own set of principles (which Google, Microsoft etc. have), which has a generally similar scope.
- The industry conversations are about companies using their own sets of principles which while similar may fall short of governments' principles in specialist areas such as education where there are unique concerns.
- Some AI principles, as currently expressed, are aspirational - they cannot yet be technically implemented or measured in whole.

[129] https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework

[130] https://www.digital.nsw.gov.au/policy/artificial-intelligence/artificial-intelligence-ethics-policy/mandatory-ethical-principles

[131] https://ministers.education.gov.au/clare/draft-schools-ai-framework-open-consultation

Enacting the principles needs participation and commitment across different organisations and jurisdictions, for example:

- Binding laws are being created in countries that don't have control over the companies developing LLMs and LFMs. And schools and systems in that county have to implement these through arrangements with those providers.
- LLM companies are joining voluntary commitments that are focused on current technical capabilities, but this could blunt motivation for more widespread ethical enforcement ambitions.
- Some AI principles extend, conflict or overlap with current policies and guidelines.
- Schools are not resourced, nor technically capable to enact the principles and will be relying on providers who are outside the jurisdiction.

This is an urgency to act given the speed of adoption of AI and the risks being experienced now and that are foreseeable, for example:

- end users are using the tools today in increasingly diverse and embedded ways,
- data are being collected today against data protection rules, and
- future AI's may be uncompetitive with those which have had the opportunity to harvest teacher and student activities to be trained from.

This means there are multiple groups, with overlapping remits, different resourcing, different jurisdiction, and there are different levels of agreement, accountabilities and even definition. It is our view that, beyond the principles, there is a need to identify activities, current and future, in pursuit of enacting the principles. The below are just some examples.

EU proposed legislation, the 'Artificial Intelligence (AI) Act'

The AI Act seeks to govern risks associated with particular practices. If, for example, AI is involved with an identified practice, then the AI can be held to account in cases of direct harm being demonstrated. The example used for education is inaccurate 'scoring of exams which impact access to education or the professional course of someone's life'.[132] This example case of a high risk means that someone who can demonstrate the AI caused harm can seek compensation from the original AI supplier. There is significant opposition with 150 Executives voicing concern[133] and a demand for a more nuanced approach which would not automatically consider all education applications of AI to be High Risk.[134]

The World Economic Forum launched its own AI Governance Alliance[135] which seeks to provide a holistic view from industry to both benefit from, and avoid the risks of, AI and technology.

It is clear that the paths for measurement and accountability of AIs are still being forged and education is a participant in developing the expectations, however, we need to be realistic that sets of principles do not create laws, nor an approach which is implementable in schools with their limited resources and competing priorities. The large models will continue to both add functionality and discover capabilities.

---

132 https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/excellence-and-trust-artificial-intelligence_en

133 https://www.ft.com/content/9b72a5f4-a6d8-41aa-95b8-c75f0bc92465

134 https://www.siia.net/wp-content/uploads/2023/02/SIIA-and-EEA-Letter-on-EU-AI-Act-9-Feb-2023.pdf

135 https://initiatives.weforum.org/ai-governance-alliance/home

Categorising the principles for action

An overarching recommendation is to examine any expressed principle through a pragmatic lens of their enactment requirements, akin to a what, how, who and when. However, even this is very difficult due to the complexity of definitions, current and future capabilities, and the implementation models available.

There are early signs of different groups trying to develop models for enacting principles. Some examples where the groups are working pragmatically are:

1. **Controlling what you can** - organisations working just within their specific remit to develop practical approaches. As noted above, the Russel group of universities have created a set of principles[136] for their staff to work with that is within the jurisdiction and control of the university:

    - Universities will support students and staff to become AI-literate.
    - Staff should be equipped to support students to use generative AI tools effectively and appropriately in their learning experience.
    - Universities will adapt teaching and assessment to incorporate the ethical use of generative AI and support equal access.
    - Universities will ensure academic rigour and integrity is upheld.
    - Universities will work collaboratively to share best practice as the technology and its application in education evolves.

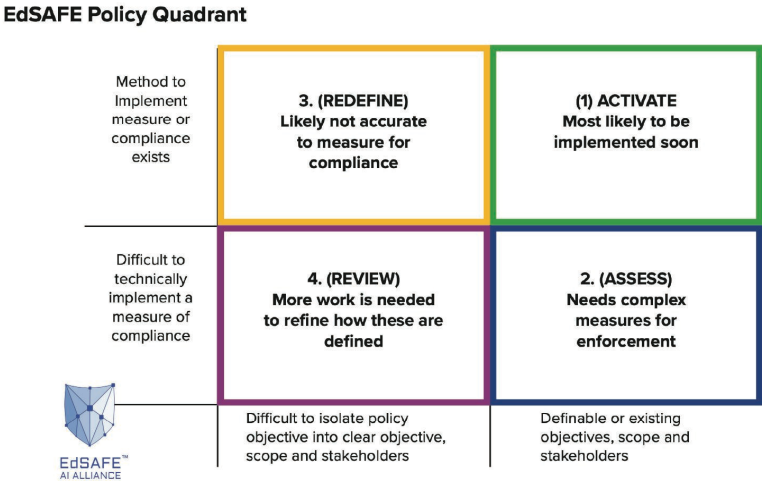
2. **Supporting what you know** - CoSN is providing data protection and advice to CIO's when enabling access to AI within schools. And AIEdu.org has role-based correspondence and toolkits[137] for supporting schools, IT leads or teachers and are developing responses to immediate and urgent use cases. For example, they are providing technical and policy responses to academic integrity issues.
3. **Influencing up** – Organisations which have government experiences and government credibility need to advocate. Digital Promise supports the office of education technology[138] working with government agencies to evolve data protection and data privacy awareness and into technical solutions that can be implemented.
4. **Guiding and informing** - developing suggested guidelines for appropriate use by teachers and practical ways that 'humans in the loop' could be maintained
5. **Sounding the alarm** - highlighting that the majority of software packages will have Gen AI capabilities within a short period of time and schools are now using it with some reporting over 50% of teachers have used[139] and therefore action is needed.

However, all of this work remains unconnected and without an organising model. There are some emerging approaches with groups like the EdSAFE AI Alliance who are developing a process for issues to be categorised into what can be actioned now, what needs further

[136] https://russellgroup.ac.uk/news/new-principles-on-use-of-ai-in-education/
[137] https://www.aiedu.org/aitoolkits
[138] https://tech.ed.gov/ai/
[139] https://www.edweek.org/technology/chatgpt-is-all-the-rage-but-teens-have-qualms-about-ai/2023/03

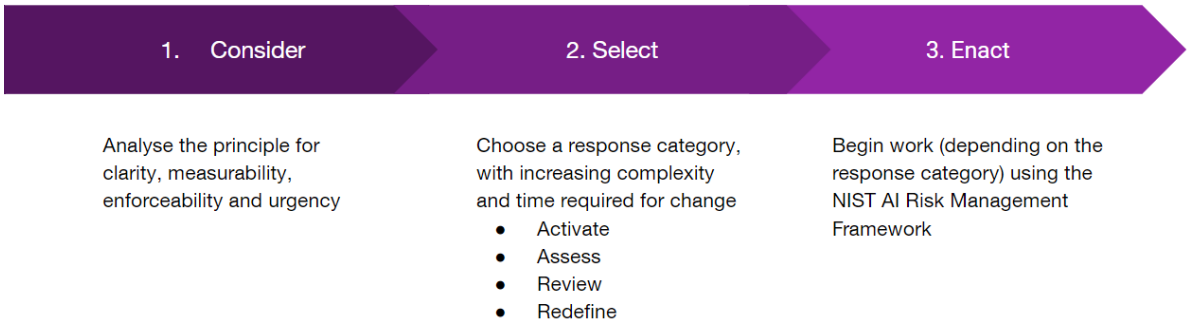
development, research or new agreements, and what needs to be reconsidered or redesigned.

The EdSAFE AI Alliance's Policy Guidance Framework provides support in determining immediate, mid-term and long-term action by implementing a quadrant prioritisation system and guiding the process from initial principle definition through to implementation and enforcement procedures.



The best practice in our view is knowing what needs to be a priority, what can be done/enacted now, who is the group who can enact and particularly focusing on your respective needs and avoid unnecessary effort on important items which will need to be addressed by other groups.

One model developed by IAMAI to support this is a 3 step Consider, Select, Enact approach which enables a practical response to any general principle for AI across different time horizons.



Consider

This stage analyses the principle under questions of:

- Clarity - Is it clear what is being asked? Is this entirely new or an amendment to something existing? Have the affected stakeholders and their changed responsibilities been identified?
- Measurability - How would the principle in action be measured? Through what means: technology audit and reporting, warrants from suppliers or oversight by a party?
- Enforceability - what jurisdiction exists over stakeholders? What enforcement powers exist now or could be developed? What other forms or influence or control could be exerted? Who would be responsible for making it happen?

Select

This stage selects a response from 4 categories of action which increase in complexity and the time required before it could be effective in schools. Using the EdSAFE policy categories of Activate, Assess, Review and Redefine with basic descriptions of:

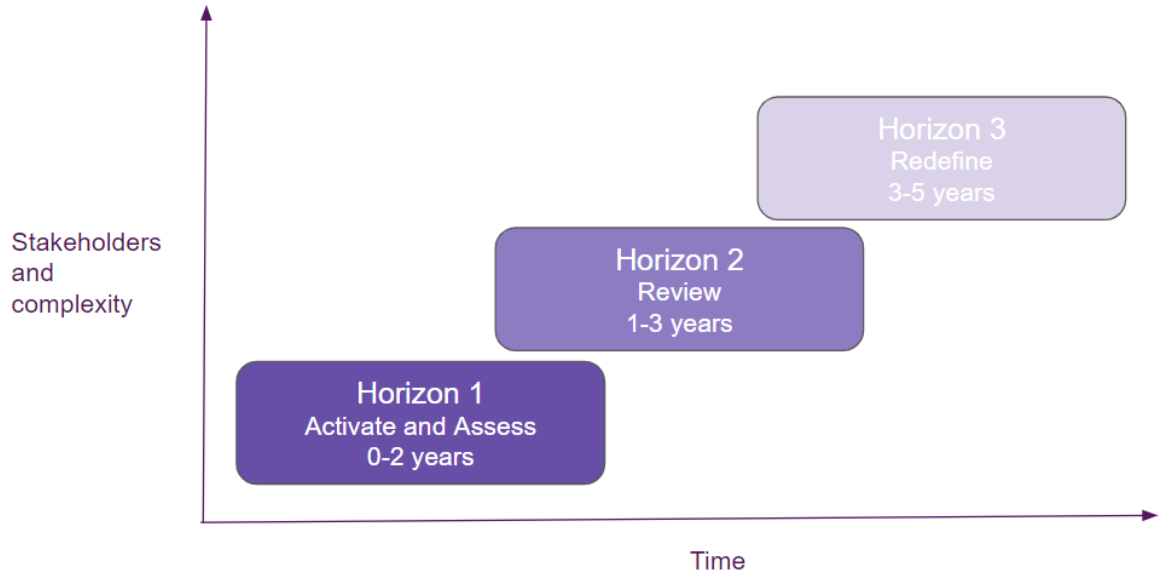
- Activate - there is a clear objective, scope of operation, measures and stakeholder control or influence. Changes required have been identified.
- Assess - there is clear objective and scope and but practical or technical issues (e.g. measurability or enforceability issues)
- Review - scope or stakeholders are not sufficiently clear and/or there are substantial practical or technical issues with enacting currently
- Redefine - there is substantial ambiguity or inability to scope or define measures and/or stakeholders in current environment

Enact

This stage guides the next step action depending on the response category. Each will draw on the NIST AI Risk Management Framework

- Activate - Plan implementation with base safety, policy change, actors and measures including pilot testing before deployment
- Assess - Work further on questions and make required changes to meet - may include controlled pilot testing
- Review - convene policy and technical research teams to explore key questions, identify and convene stakeholders, explore change in responsibilities and measures to get closer to a scope
- Redefine - Return to the principle with further exploration to determine if changes to the principle could improve the potential to enact, amend laws and regulations if required

The rough time horizons for the different response categories are shown below



Worked examples using the Draft National AI In Schools Framework

Privacy and Security Principle - **6.2 Privacy Disclosure**: students, parents, and stakeholders are proactively informed about how data will be collected, used, and shared while using a generative AI tool.

1. Consider
    a. **Clarity** - is sufficient, but some updates would improve it: the principle is clear on the definition of scope of users to be included, it is clear regarding the intent of target context being 'Gen AI', however, that needs a nomenclature to bind it to the context of 'while using'. It is somewhat clear via an inference on what would be included by term 'data'. The terms 'collected', 'user' and 'shared' should be aligned with existing data protection scoping to support the intended inference and to provide a complete spectrum of the term 'informed'. There is reasonable clarity about the required action of informed consent, however, it is not clear what 'proactively informed' implies leaving scope interpretation e.g. is a privacy statement in terms and conditions sufficient, and who has authority to acknowledge and make these consents?
    b. **Measurability** - practical measures exist. We can infer the expectation is on the GenAI provider to undertake the measure, however, measurement oversight with consumer-based systems has hurdles to overcome. Actions to measure compliance with consultation advice and consent from the gen AI provider through interfaces and prompts recorded and an audit trail can be provided for accountability.
    c. **Enforceability** – It is clear that action could be taken against gen AI providers gathering the informed consent and clear obligation and ability for school or system IT. The scope for enforcement needs to consider where Gen AI tools are composite tools within other systems (e.g. a chat bot add-on to a site), when they are not licenced to an educational institution, and when they are governed and hosted by an educational authority. The enforcements are significantly different for each of these implementations.
    d. **Urgency** – There is a strong social need now with use from these stakeholders without guiding advice from anyone other than the gen AI provider
2. Select – on the basis of the Consider stage we can select 'Activate'
3. Enact
    a. Ensure Base Safety in schools with authentication and tracking for users
    b. Updates to system and school policies on gen AI advising disclosure is needed
    c. Gen AI vendor engagement to present the disclosure

Human and Social Wellbeing - **2.2 Diversity of Perspectives**: generative AI tools expose users to diverse ideas and perspectives, avoiding the reinforcement of existing biases.

1. Consider

    a. Clarity - the principle is clear on the objective but there is ambiguity about 'exposing' users, e.g. what constitutes diverse without breaching non-discrimination principle and also meeting different school values and which existing biases?
    b. Measurability - significant challenge to measure level of exposure or the ideas and perspectives given qualitative assessments.
    c. Enforceability - difficult to determine if applied to LLM provider or others in delivering to system or school. If LLM jurisdiction and control issues. But likely LLMs will be developing ways to manage over the medium term
    d. Urgency - not immediate safety issues but increasing potential for corrosion if not addressed
2. Select - on the basis of the Consider stage we can select Review
3. Enact
    a. Convene working groups to define key terms and ways of measuring and understanding adherence to the principle and ways to manage specific school values and ethos
    b. Work with other Australian framework providers (e.g., Industry) to determine best way to engage with LLM providers to understand technical potential or roadmaps for giving effect to principles
    c. Undertake controlled testing of LLMs to determine specific areas of concern for targeted response

Appendix 1: A selection of ethically focused AI principles & taxonomies

Australian Government AI Ethics Principles

https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles

OECD AI Principles overview

https://oecd.ai/en/ai-principles

OECD AI Classification Framework

https://oecd.ai/en/classification

The Ethical Framework for AI in Education

https://www.buckingham.ac.uk/wp-content/uploads/2021/03/The-Institute-for-Ethical-AI-in-Education-The-Ethical-Framework-for-AI-in-Education.pdf

Ethical principles for artificial intelligence in K-12 education

https://doi.org/10.1016/j.caeai.2023.100131

Berkman Kline

https://cyber.harvard.edu/publication/2020/principled-ai

Privacy: • Privacy • Control over Use of Data • Consent • Privacy by Design • Recommendation for Data Protection Laws Ability to Restrict Processing • Right to Rectification • Right to Erasure	Accountability: • Accountability • Recommendation for New Regulations • Impact Assessment • Evaluation and Auditing Requirement • Verifiability and Replicability • Liability and Legal Responsibility • Ability to Appeal • Environmental Responsibility • Creation of a Monitoring Body Remedy for Automated Decision	Transparency and Explainability: • Explainability • Transparency • Open Source Data and Algorithms • Notification when Interacting with an AI • Notification when AI Makes a Decision about an Individual Regular Reporting Requirement • Right to Information • Open Procurement (for Government)	Fairness and Non-discrimination: • Non-discrimination and the Prevention of Bias Fairness • Inclusiveness in Design • Inclusiveness in Impact • Representative and High Quality Data • Equality
Safety and Security: • Security • Safety and Reliability • Predictability • Security by Design	Professional Responsibility: • Multi Stakeholder Collaboration Responsible Design • Consideration of Long Term Effects Accuracy	Human Control of Technology: • Human Control of Technology • Human Review of Automated Decision • Ability to Opt out of	Promotion of Human Values: • Leveraged to Benefit Society • Human Values and Human Flourishing

	• Scientific Integrity	Automated Decision	

Australia's artificial intelligence ethics framework

Australian Government AI Ethics Principles

https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles

Australian Government Principles at a glance

- Human, societal and environmental wellbeing: AI systems should benefit individuals, society and the environment.
- Human-centred values: AI systems should respect human rights, diversity, and the autonomy of individuals.
- Fairness: AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups.
- Privacy protection and security: AI systems should respect and uphold privacy rights and data protection, and ensure the security of data.
- Reliability and safety: AI systems should reliably operate in accordance with their intended purpose.
- Transparency and explainability: There should be transparency and responsible disclosure so people can understand when they are being significantly impacted by AI, and can find out when an AI system is engaging with them.
- Contestability: When an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or outcomes of the AI system.
- Accountability: People responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and human oversight of AI systems should be enabled.

Microsoft responsible AI principles

https://www.microsoft.com/en-us/ai/our-approach

- Fairness
- AI systems should treat all people fairly
- Reliability & Safety
- AI systems should perform reliably and safely
- Privacy & Security
- AI systems should be secure and respect privacy
- Inclusiveness
- AI systems should empower everyone and engage people

- Transparency
- AI systems should be understandable
- Accountability
- People should be accountable for AI systems

IBM Ethical AI Principles:

IBM Ethical AI Principles:
The following represents six ethical AI principles of IBM:

- Accountability: AI designers and developers are responsible for considering AI design, development, decision processes, and outcomes.
- Value alignment: AI should be designed to align with the norms and values of your user group in mind.
- Explainability: AI should be designed for humans to easily perceive, detect, and understand its decision process, and the predictions/recommendations. This is also, at times, referred to as interpretability of AI. Simply speaking, users have all rights to ask the details on the predictions made by AI models such as which features contributed to the predictions by what extent. Each of the predictions made by AI models should be able to be reviewed.
- Fairness: AI must be designed to minimize bias and promote inclusive representation.
- User data rights: AI must be designed to protect user data and preserve the user's power over access and uses.

OECD AI Principles overview

https://oecd.ai/en/ai-principles

Values-based principles

Inclusive growth, sustainable development and well-being
Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, sustainable development and well-being.

Human-centred values and fairness

AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognised labour rights.
To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art.

Rationale

AI should be developed consistent with human-centred values, such as fundamental freedoms, equality, fairness, rule of law, social justice, data protection and privacy, as well as consumer rights and commercial fairness.

Some applications or uses of AI systems have implications for human rights, including risks that human rights (as defined in the Universal Declaration of Human Rights)1 and human-centred values might be deliberately or accidentally infringed. It is therefore important to promote "values-alignment" in AI systems (i.e., their design with appropriate safeguards) including capacity for human intervention and oversight, as appropriate to the context. This alignment can help ensure that AI systems' behaviours protect and promote human rights and align with human-centred values throughout their operation. Remaining true to shared democratic values will help strengthen public trust in AI and support the use of AI to protect human rights and reduce discrimination or other unfair and/or unequal outcomes.

This principle also acknowledges the role of measures such as human rights impact assessments (HRIAs) and human rights due diligence, human determination (i.e., a "human in the loop"), codes of ethical conduct, or quality labels and certifications intended to promote human-centred values and fairness.

Transparency and explainability

AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art:

- to foster a general understanding of AI systems,
- to make stakeholders aware of their interactions with AI systems, including in the workplace,
- to enable those affected by an AI system to understand the outcome, and,
- to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.

Rationale for this principle

The term transparency carries multiple meanings. In the context of this Principle, the focus is first on disclosing when AI is being used (in a prediction, recommendation or decision, or that the user is interacting directly with an AI-powered agent, such as a chatbot). Disclosure should be made with proportion to the importance of the interaction. The growing ubiquity of AI applications may influence the desirability, effectiveness or feasibility of disclosure in some cases.

Transparency further means enabling people to understand how an AI system is developed, trained, operates, and deployed in the relevant application domain, so that consumers, for example, can make more informed choices. Transparency also refers to the ability to provide meaningful information and clarity about what information is provided and why. Thus

transparency does not in general extend to the disclosure of the source or other proprietary code or sharing of proprietary datasets, all of which may be too technically complex to be feasible or useful to understanding an outcome. Source code and datasets may also be subject to intellectual property, including trade secrets.

An additional aspect of transparency concerns facilitating public, multi-stakeholder discourse and the establishment of dedicated entities, as necessary, to foster general awareness and understanding of AI systems and increase acceptance and trust.

Explainability means enabling people affected by the outcome of an AI system to understand how it was arrived at. This entails providing easy-to-understand information to people affected by an AI system's outcome that can enable those adversely affected to challenge the outcome, notably – to the extent practicable – the factors and logic that led to an outcome. Notwithstanding, explainability can be achieved in different ways depending on the context (such as, the significance of the outcomes). For example, for some types of AI systems, requiring explainability may negatively affect the accuracy and performance of the system (as it may require reducing the solution variables to a set small enough that humans can understand, which could be suboptimal in complex, high-dimensional problems), or privacy and security. It may also increase complexity and costs, potentially putting AI actors that are SMEs at a disproportionate disadvantage.

Therefore, when AI actors provide an explanation of an outcome, they may consider providing – in clear and simple terms, and as appropriate to the context – the main factors in a decision, the determinant factors, the data, logic or algorithm behind the specific outcome, or explaining why similar-looking circumstances generated a different outcome. This should be done in a way that allows individuals to understand and challenge the outcome while respecting personal data protection obligations, if relevant.

Robustness, security and safety

AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk.
To this end, AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system's outcomes and responses to inquiry, appropriate to the context and consistent with the state of art.
AI actors should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias.

Rationale for this principle
Addressing the safety and security challenges of complex AI systems is critical to fostering trust in AI. In this context, robustness signifies the ability to withstand or overcome adverse conditions, including digital security risks. This principle further states that AI systems should not pose unreasonable safety risks including to physical security, in conditions of normal or

foreseeable use or misuse throughout their lifecycle. Existing laws and regulations in areas such as consumer protection already identify what constitutes unreasonable safety risks. Governments, in consultation with stakeholders, must determine to what extent they apply to AI systems.

AI actors can employ a risk management approach (see below) to identify and protect against foreseeable misuse, as well as against risks associated with use of AI systems for purposes other than those for which they were originally designed. Issues of robustness, security and safety of AI are interlinked. For example, digital security can affect the safety of connected products such as automobiles and home appliances if risks are not appropriately managed.

The Recommendation highlights two ways to maintain robust, safe and secure AI systems: traceability and subsequent analysis and inquiry, and applying a risk management approach. like explainability (see 1.3), traceability can help analysis and inquiry into the outcomes of an AI system and is a way to promote accountability. Traceability differs from explainability in that the focus is on maintaining records of data characteristics, such as metadata, data sources and data cleaning, but not necessarily the data themselves. In this, traceability can help to understand outcomes, to prevent future mistakes, and to improve the trustworthiness of the AI system.

Risk management approach: The Recommendation recognises the potential risks that AI systems pose to human rights, bodily integrity, privacy, fairness, equality and robustness. It further recognises the costs of protecting from these risks, including by building transparency, accountability, safety and security into AI systems. It also recognises that different uses of AI present different risks, and some risks require a higher standard of prevention or mitigation than others.

A risk management approach, applied throughout the AI system lifecycle, can help to identify, assess, prioritise and mitigate potential risks that can adversely affect a system's behaviour and outcomes. Other OECD standards on risk management, for example in the context of digital security risk management and risk-based due diligence under the MNE Guidelines and OECD Due Diligence Guidance for Responsible Business Conduct, may offer useful guidance1. Documenting risk management decisions made at each lifecycle phase can contribute to the implementation of the other principles of transparency (1.3) and accountability (1.5).

Accountability

AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art.

Rationale for this principle

The terms accountability, responsibility and liability are closely related yet different, and also carry different meanings across cultures and languages. Generally speaking, "accountability" implies an ethical, moral, or other expectation (e.g., as set out in management practices or codes of conduct) that guides individuals' or organisations' actions or conduct and allows

them to explain reasons for which decisions and actions were taken. In the case of a negative outcome, it also implies taking action to ensure a better outcome in the future. “Liability” generally refers to adverse legal implications arising from a person’s (or an organisation’s) actions or inaction. “Responsibility” can also have ethical or moral expectations and can be used in both legal and non-legal contexts to refer to a causal link between an actor and an outcome.

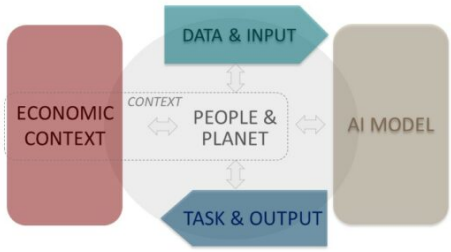
Given these meanings, the term “accountability” best captures the essence of this principle. In this context, “accountability” refers to the expectation that organisations or individuals will ensure the proper functioning, throughout their lifecycle, of the AI systems that they design, develop, operate or deploy, in accordance with their roles and applicable regulatory frameworks, and for demonstrating this through their actions and decision-making process (for example, by providing documentation on key decisions throughout the AI system lifecycle or conducting or allowing auditing where justified).

OECD AI Classification Framework

https://oecd.ai/en/classification

The framework allows users to zoom in on specific risks that are typical of AI, such as bias, explainability and robustness, yet it is generic in nature. It facilitates nuanced and precise policy debate. The framework can also help develop policies and regulations, since AI system characteristics influence the technical and procedural measures they need for implementation. In particular, the framework provides a baseline framework to help support and advance:

- A common understanding of AI, and metrics
- Registries or inventories of AI systems
- Sector-specific frameworks, e.g. in healthcare
- Risk assessment, incident reporting and risk management



Classification framework dimensions and criteria at a glance

PEOPLE & PLANET	Criteria	Description
USERS	Users of AI system	What is the level of competency of users who interact with the system?
STAKEHOLDERS	Impacted stakeholders	Who is impacted by the system (e.g. consumers, workers, government agencies)?
OPTIONALITY	Optionality and redress	Can users opt out, e.g. switch systems? Can users challenge or correct the output?
HUMAN RIGHTS	Human rights and democratic values	Can the system's outputs impact fundamental human rights (e.g. human dignity, privacy, freedom of expression, non-discrimination, fair trial, remedy, safety)?
WELL-BEING & ENVIRONMENT	Well-being, society and the environment	Can the system's outputs impact areas of life related to well-being (e.g. job quality, the environment, health, social interactions, civic engagement, education)?
DISPLACEMENT	*{Displacement potential}*	*Could the system automate tasks that are or were being executed by humans?*
ECONOMIC CONTEXT	**Criteria**	**Description**
SECTOR	Industrial sector	Which industrial sector is the system deployed in (e.g. finance, agriculture)?
BUSINESS FUNCTION & MODEL	Business function	What business function(s) is the system employed in (e.g. sales, customer service)?
	Business model	Is the system a for-profit use, non-profit use or public service system?
CRITICALITY	Impacts critical functions / activities	Would a disruption of the system's function / activity affect essential services?
SCALE & MATURITY	Breadth of deployment	Is the AI system deployment a pilot, narrow, broad or widespread?
	{Technical maturity}	*How technically mature is the system (Technology Readiness Level –TRL)*
DATA & INPUT	**Criteria**	**Description**
COLLECTION	Detection and collection	Are the data and input collected by humans, automated sensors or both?
	Provenance of data and input	Are the data and input from experts; provided, observed, synthetic or derived?
	Dynamic nature	Are the data dynamic, static, dynamic updated from time to time or real-time?
RIGHTS & IDENTIFIABILITY	Rights	Are the data proprietary, public or personal data (related to identifiable individual)?
	"Identifiability" of personal data	If personal data, are they anonymised; pseudonymised?
STRUCTURE & FORMAT	*{Structure of data and input}*	*Are the data structured, semi-structured, complex structured or unstructured?*
	{Format of data and metadata}	*Is the format of the data and metadata standardised or non-standardised?*
SCALE	*{Scale}*	*What is the dataset's scale?*
QUALITY AND APPROPRIATENESS	*{Data quality and appropriateness}*	*Is the dataset fit for purpose? Is the sample size adequate? Is it representative and complete enough? How noisy are the data?*
AI MODEL	**Criteria**	**Description**
MODEL CHARACTERISTICS	Model information availability	Is any information available about the system's model?
	AI model type	Is the model symbolic (human-generated rules), statistical (uses data) or hybrid?
	{Rights associated with model}	*Is the model open-source or proprietary, self or third-party managed?*
	{Discriminative or generative}	*Is the model generative, discriminative or both?*
	{Single or multiple model(s)}	*Is the system composed of one model or several interlinked models?*
MODEL-BUILDING	Model-building from machine or human knowledge	Does the system learn based on human-written rules, from data, through supervised learning, through reinforcement learning?
	Model evolution in the field ML	Does the model evolve and / or acquire abilities from interacting with data in the field?
	Central or federated learning ML	Is the model trained centrally or in a number of local servers or "edge" devices?
MODEL INFERENCE	*{Model development / maintenance}*	*Is the model universal, customisable or tailored to the AI actor's data?*
	{Deterministic and probabilistic}	*Is the model used in a deterministic or probabilistic manner?*
	Transparency and explainability	If information available to users to allow them to understand model outputs?
TASK & OUTPUT	**Criteria**	**Description**
TASKS	Task(s) of the system	What tasks does the system perform (e.g. recognition, event detection, forecasting)?
	{Combining tasks and actions into composite systems}	*Does the system combine several tasks and actions (e.g. content generation systems, autonomous systems, control systems)?*
ACTION	Action autonomy	How autonomous are the system's actions and what role do humans play?
APPLICATION AREA	Core application area(s)	Does the system belong to a core application area such as human language technologies, computer vision, automation and / or optimisation or robotics?
EVALUATION	*{Evaluation methods}*	*Are standards or methods available for evaluating system output?*

Note: Criteria and descriptions in grey and marked with an {} symbol = those where objective and consistent information is available. ML = for machine learning AI models.

More information: www.oecd.ai/classification | Contact: ai@oecd.org

Appendix 2: Set of general data protection taxonomies

GDPR

https://edps.europa.eu/data-protection/data-protection/legislation_en

1. Lawfulness, fairness, and transparency
Whenever you're processing personal data, you should have a good reason for doing so. GDPR terms this principle lawfulness. Reasons for processing data can include:

- The user has given you consent to do so.
- You must do it to make good on a contract.
- It's necessary to fulfill a legal obligation.
- For protection of vital interests of a natural person.
- It's a public task done in public interest.
- You can prove you have legitimate interest, and it's not overridden by data subject's rights and interests.

The concept of fairness laid out in the GDPR goes hand-in-hand with lawfulness. It means you shouldn't purposely withhold information about what or why you're collecting data. In other words, users wouldn't be surprised if they knew how you were using their data. Fairness means you won't mishandle or misuse the data you collect.

Transparency is inherently linked to fairness: Being clear, open, and honest with data subjects about who you are, and why and how you're processing their personal data is the definition of transparency. By following it, you act fairly towards your data subjects.

2. Purpose limitation
The GDPR's second principle sets boundaries around using data only for specific activities. This purpose limitation means data is "collected for specified, explicit, and legitimate purposes" only, as stated in the GDPR.

Your purposes for processing data must be clearly established. And they must also be clearly communicated to individuals through a privacy notice. Finally, you must follow them closely, limiting the processing of data to only the purposes you've stated.

If at any point, you want to use the data you've collected for a new purpose that's incompatible with your original purpose, you must ask specifically for consent again to do it — unless you have a clear obligation or function set out in law.

3. Data minimization
Only collect the smallest amount of data you'll need to complete your purposes. This is the GDPR principle of data minimization. For example, if you want to gather subscribers for your email newsletter, you should only ask for information necessary to send out the newsletters. Avoid gathering personal data such as phone numbers or home addresses, which aren't directly related to your purpose.

4. Accuracy

It's up to you to ensure the accuracy of the data you collect and store. Set up checks and balances to correct, update, or erase incorrect or incomplete data that comes in. Also have regular audits on the calendar to double-check the cleanliness of stored data.

5. Storage limitation
According to the GDPR, you have to justify the length of time you're keeping each piece of data you store. Data retention periods are a good thing to establish to meet this storage limitation policy. Create a standard time period after which you'll anonymize any data you're not actively using.

6. Integrity and confidentiality
The GDPR requires you maintain the integrity and confidentiality of the data you collect, essentially keeping it secure from internal or external threats. This takes planning and proactive diligence. You must protect data from unauthorized or unlawful processing and accidental loss, destruction, or damage.

7. Accountability
The GDPR regulators know an organization can say they're following all the rules without actually doing it. That's why they require a level of accountability: You must have appropriate measures and records in place as proof of your compliance with the data processing principles. Supervisory authorities can ask for this evidence at any time. Documentation is key here. It creates an audit trail you — and authorities — can follow if you do need to prove responsibility.

Conclusion: Integration of the 7 Principles of the GDPR
The 7 principles of the GDPR communicate the spirit and thought process behind data processing best practices. In addition, the GDPR sets out data controller and processor responsibilities that support each of the principles.

Instead of being a piece of the operational puzzle, these 7 principles inform all processing activity and business practices — from the design stage across the entire data processing lifecycle. This can be best fulfilled by implementing privacy by design and default.

Safer Technology for schools

Details of each Criteria are located in the linked PDF document.
https://st4s.edu.au/wp-content/uploads/2023/04/Safer-Technologies-4-Schools-Supplier-Guide-2022.1-v1.03.pdf

6. Assessment Criteria
Criteria – Security
6.2.1 Security – Product function
6.2.2 Security – Hosting and Location.
6.2.3 Security – Technical
6.2.5 Security – Access
Security – Processes and Testing
6.2.8 Security – Plans and Quality Security – Incidents
6.2.10 Security – Data Deletion and Retention

6.2.11 Security – Compliance Controls
6.2.12 Security – Governance

6.3 Criteria – Privacy Privacy
6.3.2 Privacy – Requests
6.3.3 Privacy – Functionality

6.4 Criteria Interoperability
6.4.1 Interoperability – Data Standards.
6.4.2 Interoperability – Technical Integration
6.4.3 Interoperability – Data Availability

US based Data protection and privacy guides

Appendix 3: Set of educational taxonomies

Below are some examples of educational focused taxonomies

Framework for Improving Student Outcomes (FISO)

https://www2.education.vic.gov.au/pal/fiso/policy

Learning
Learning is the ongoing acquisition by students of knowledge, skills and capabilities,.

Wellbeing
Wellbeing is the development of the capabilities necessary to thrive, contribute and respond positively to challenges and opportunities of life.

Leadership	The strategic direction and deployment of resources to create and reflect shared goals and values; high expectations; and a positive, safe and orderly learning environment
	Shared development of a culture of respect and collaboration with positive and supportive relationships between students and staff at the core
Teaching and Learning	Documented teaching and learning program based on the Victorian Curriculum and senior secondary pathways, incorporating extra-curricula programs
	Use of common and subject-specific high impact teaching and learning strategies as part of a shared and responsive teaching and learning model implemented through positive and supportive student-staff relationships
Assessment	Systematic use of assessment strategies and measurement practices to obtain and provide feedback on student learning growth, attainment and wellbeing capabilities
	Systematic use of data and evidence to drive the prioritisation, development, and implementation of actions in schools and classrooms
Engagement	Activation of student voice and agency, including in leadership and learning, to strengthen students' participation and engagement in school
	Strong relationships and active partnerships between schools and families/carers, communities, and organisations to strengthen students' participation and engagement in school
Support and resources	Responsive, tiered and contextualised approaches and strong relationships to support student learning, wellbeing and inclusion
	Effective use of resources and active partnerships with families/carers, specialist providers and community organisations to provide responsive support to students

COSN - Education Response to Artificial Intelligence & Generative AI

Essential Leadership Guidelines
Generative AI has ushered in a paradigm shift in society that K-12 institutions can shepherd. Superintendents and school district administrators are encouraged to implement these essential guidelines when working with leadership teams and staff to create actionable steps and policies around AI and Generative AI.
Awareness: Ensure that users are aware of the AI tools and their potential benefits for K-12 education. Focusing on how to use Generative AI as a way to develop higher-order thinking skills is a good start.
Limitations: Explain the limitations of the AI tools and the potential for errors or inaccuracies. Teach critical thinking skills to assess and validate AI output.
Ethics and Etiquette: Promote good online etiquette, including proofreading and fact-checking. Teach Ethics in
relationship to AI created or assisted work products.
Ongoing Training: Provide ongoing innovation training and reinforcement on the best ways to use AI tools in a
safe and responsible manner.
Reporting: Educate the school community about how to report incidents or concerns.
Policies: Set policies to create a culture of safe and responsible use to mitigate the potential risks associated with
using AI tools in a school environment while iterating effective ways to leverage the power of generative AI.
Privacy and Security Measures: Review your student data privacy policy, practices, and security measures and
consider how they relate when using AI tools

Education for AI, not AI for Education: The Role of Education and Ethics in National AI Policy Strategies

https://link.springer.com/article/10.1007/s40593-021-00270-2

Education for AI (i.e., training)
• Training AI Experts: discussion of developing future AI practitioners, such as computer scientists and engineers.
• Preparing the Workforce for AI: discussion of education and training efforts to help workers adapt to labor disruption due to AI.
• Public AI Literacy: discussion of the need to educate the broader public about AI.
AI for Education (i.e., AIED)
• •
Teaching and Learning: discussion of AI-based teaching and learning tools such as intelligent tutoring systems, pedagogical agents, and predictive assessments. Administrative Tools: discussion of AI used to support administration in educa- tional systems, for example, to make admission, promotion, or graduation deci- sions.

The Ethical Framework for AI in Education

https://www.buckingham.ac.uk/wp-content/uploads/2021/03/The-Institute-for-Ethical-AI-in-Education-The-Ethical-Framework-for-AI-in-Education.pdf

Achieving Educational Goals. AI should be used to achieve well-defined educational goals based on strong societal, educational or scientific evidence that this is for the benefit of learner	
AI should be used to assess and recognise a broader range of learners' talents	
AI should increase the capacity of educational institutions whilst respecting human Relationships	
AI systems should be used in ways that promote equity between different groups of learners and not in ways that discriminate against any group of learners	
AI should be used to increase the level of control that learners have over their learning and development	
A balance should be struck between privacy and the legitimate use of data for achieving well-defined and desirable educational goals	
Humans are ultimately responsible for educational outcomes and should therefore have an appropriate level of oversight of how AI systems operate	
Learners and educators should have a reasonable understanding of artificial intelligence and its implications	
.AI resources should be designed by people who understand the impacts these resources will have	

https://www.buckingham.ac.uk/wp-content/uploads/2021/03/The-Institute-for-Ethical-AI-in-Education-Annex-Developing-the-Ethical-Framework-for-AI-in-Education-IEAIED-.pdf

Ethical principles for artificial intelligence in K-12 education

https://doi.org/10.1016/j.caeai.2023.100131
Summary of 4 educational aligned ethical principles

Analysis of AIEdK-12 ethics guidance documents according to ethical principle.

AI Ethics Guideline Document ->	World Economic Forum (2019)	IEAIED (2021a)	UNESCO (2021)	UNICEF (2021)
Ethical Principle/Constituency	International	UK & beyond	International	International
Transparency[a] Key words: Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing, *age-appropriate language*	Ensuring algorithmic accountability	***Transparency and Accountability***	(addressed under "***Overarching principle for AI***")	***Provide transparency, explainability, and accountability for children***
Justice & fairness[a] Key words: Justice, fairness, consistency, inclusion, equality, equity, (non-)bias, (non-) discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution	Accounting for marginalized groups; Ensuring fairness in machine learning	***Equity***	***Policies and regulations for equitable, inclusive, and ethical use of AI***	***Ensure inclusion of and for children; Prioritize fairness & non-discrimination for children***
Non-maleficence[a] Key words: Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion	(addressed under other categories)	***Ethical Design***	(addressed under other categories)	***Ensure safety for children***
Responsibility[a] Key words: Responsibility, accountability, liability, acting with integrity	Consumer Protection	***Transparency and Accountability***	(addressed under "***Overarching principle for AI***")	***Provide transparency, explainability, and accountability for children***
Privacy[a] Key words: Privacy, personal or private information; *"right to be forgotten" (RTBF)*	Privacy	***Privacy***	Establish data protection laws which make educational data collection and analysis visible, traceable, and auditable by teachers, students and parents	***Protect children's data and privacy***
Beneficence[a] Key words: Benefits, beneficence, well-being, peace, social good, common good	Recognizing developmental science in policy	***Achieving Educational Goals***	(addressed under other categories: "Emphasize students' agency and social well-being in the process of integrating AI-based tools")	***Support children's development and well-being***
Freedom & autonomy[a] Key words: Freedom, autonomy, consent, *assent*, choice, self-determination, liberty, empowerment	Agency	***Autonomy; Informed Participation***	Cultivate learner-centred use of AI to enhnce learning and assessment (reinforce and reiterate human's autority and autonomy over their own learning)	***Protect children's data and privacy; Ensure safety for children***
Pedagogical appropriateness Keywords: Appropriate use, educational research-based, evidence-based, alignment with learner needs, child-centred AI, developmentally appropriate	Algorithms for Children; Assessment and Evaluation	***Achieving Educational Goals; Forms of Assessment***	***Pilot testing, monitoring and evaluation, and building an evidence base***	***Create an enabling environment; Support children's development and well-being***
Children's rights Keywords: Children's or child rights	***Child Rights***	(addressed under other categories)	(addressed under "***Overarching principle for AI***")	***Empower governments and businesses with knowledge of AI and children's rights***
AI literacy Keywords: AI literacy; digital literacy; AI curriculum; AI education	Public education	***Informed Participation***	Integrate AI-related skills into school curricula	***Prepare children for present and future developments in AI***
Teachers' well-being Keywords: Teacher well-being; teacher workload; teacher empowerment	–	***Administration and Workload***	Ensure that AI is used to empower teachers	–

Screen shot from summary table