OECD Education Working Papers No. 218

Trustworthy artificial intelligence (AI) in education: Promises and challenges

Stéphan Vincent-Lancrin,

Reyer van der Vlies

https://dx.doi.org/10.1787/a6c90fa9-en

OECD

Organisation for Economic Co-operation and Development

EDU/WKP(2020)6

Unclassified **English text only**

6 April 2020

DIRECTORATE FOR EDUCATION AND SKILLS

Trustworthy artificial intelligence (AI) in education: promises and challenges

OECD Education Working Paper No. 218

By Stéphan Vincent-Lancrin (OECD) and Reyer van der Vlies (OECD)

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

Stéphan Vincent-Lancrin (Stephan.vincent-lancrin@oecd.org)
Reyer van der Vlies (Reyer.vandervlies@oecd.org)

JT03460412

OECD EDUCATION WORKING PAPERS SERIES

OECD Working Papers should not be reported as representing the official views of the OECD or of its member countries. The opinions expressed and arguments employed herein are those of the author(s).

Working Papers describe preliminary results or research in progress by the author(s) and are published to stimulate discussion on a broad range of issues on which the OECD works. Comments on Working Papers are welcome, and may be sent to the Directorate for Education and Skills, OECD, 2 rue André-Pascal, 75775 Paris Cedex 16, France.

Comment on the series is welcome, and should be sent to edu.contact@oecd.org.

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

---

www.oecd.org/edu/workingpapers

Abstract

This paper was written to support the G20 artificial intelligence (AI) dialogue. With the rise of artificial intelligence (AI), education faces two challenges: reaping the benefits of AI to improve education processes, both in the classroom and at the system level; and preparing students for new skillsets for increasingly automated economies and societies. AI applications are often still nascent, but there are many examples of promising uses that foreshadow how AI might transform education. With regard to the classroom, this paper highlights how AI can accelerate personalised learning, the support of students with special needs. At the system level, promising uses include predictive analysis to reduce dropout, and assessing new skillsets. A new demand for complex skills that are less easy to automate (e.g. higher cognitive skills like creativity and critical thinking) is also the consequence of AI and digitalisation. Reaching the full potential of AI requires that stakeholders trust not only the technology, but also its use by humans. This raises new policy challenges around "trustworthy AI", encompassing the privacy and security of data, but also possible wrongful uses of data leading to biases against individuals or groups.

Résumé

Ce document a été rédigé pour soutenir le dialogue du G20 sur l'intelligence artificielle (IA). Avec l'essor de l'intelligence artificielle (IA), l'éducation est confrontée à deux défis : récolter les bénéfices de l'IA pour améliorer les processus éducatifs, tant dans les salles de classe qu'au niveau du système, et préparer les étudiants à de nouveaux ensembles de compétences pour des économies et des sociétés de plus en plus automatisées. Les applications de l'IA sont souvent encore naissantes, mais il existe de nombreux exemples d'utilisations prometteuses qui laissent entrevoir comment l'IA pourrait transformer l'éducation. En ce qui concerne la salle de classe, ce document souligne comment l'IA peut accélérer l'apprentissage personnalisé ou le soutien des étudiants ayant des besoins particuliers. Au niveau du système, les utilisations prometteuses comprennent l'analyse prédictive pour réduire les abandons scolaires ou l'évaluation de nouvelles compétences. Une nouvelle demande de compétences complexes et moins faciles à automatiser (par exemple, des compétences cognitives plus élevées comme la créativité et la pensée critique) est également la conséquence de l'IA et de la numérisation. Pour atteindre le plein potentiel de l'IA, il faut que les parties prenantes fassent confiance non seulement à la technologie, mais aussi à son utilisation par les humains, ce qui soulève de nouveaux défis politiques autour de l'"IA digne de confiance". Cela inclut une attention à la confidentialité et la sécurité des données, mais aussi aux éventuelles utilisations abusives des données conduisant à des préjugés contre des individus ou des groupes.

Table of contents

Abstract 3

Résumé 3

Foreword 5

1. Introduction 6

2. Artificial intelligence in education 6
   2.1. AI applications for instruction 7
   2.2. AI applications for system and school management 9

3. Skills for the digital era 11

4. Policy challenges 12
   4.1. Trust in AI 14
   4.2. Dealing with privacy and security 15

5. Conclusions 16

References 17

Boxes

Box 2.1. Artificial intelligence 7
Box 4.1. G20 AI Principles 13

Foreword

In 2019 in Osaka, Japan, G20 Leaders welcomed the G20 Artificial Intelligence (AI) Principles, committing to a human-centred approach to AI in order to foster public trust and confidence in AI technologies and fully realise their potential. As elaborated in the 2019 G20 Ministerial Statement on Trade and Digital Economy, the G20 AI Principles for responsible stewardship of trustworthy AI put a spotlight on inclusive growth, sustainable development and well-being; human-centred values and fairness; transparency and explainability; robustness, security and safety; and accountability. The Principles stem from the recognition that while AI technologies have the potential to help advance the sustainable development goals (SDGs) and realise a sustainable and inclusive society, they also may present societal challenges, including labour market transitions, privacy, security, ethical issues, new digital divides and the need for AI capacity building.

Following the commitment of G20 Leaders, and to strengthen the G20's path setting role in promoting the responsible development and use of AI, the 2020 Saudi Arabian G20 Presidency has proposed to establish a G20 AI Dialogue. This Dialogue, to take place under the G20 Digital Economy Task Force (DETF), has the broad aims to keep global policy making on AI up to speed with technological developments, to support sharing of experiences and policy practices amongst G20 members, and to inform the G20's ongoing implementation of the G20 AI Principles. It complements the Saudi Presidency's parallel efforts in 2020 to develop a G20 Action Plan on implementation of the G20 AI Principles.

In 2020, the G20 AI Dialogue focused on trustworthy AI in education, healthcare and public services (e-government). Against the backdrop of the G20 AI Principles, it aims to discuss how AI applications at the sector level uphold responsible stewardship of trustworthy AI and the challenges that arise as these sectors make increasing use of this technology.

This working paper was developed as a background paper to support the first G20 AI Dialogue discussion in 2020, on "Trustworthy AI in Education". It aims to provide with a selected array of AI use cases in education and a glimpse into the opportunities and challenges raised by this technology, with a view to stimulating debate and providing sector-level insights into how the G20 AI Principles can be successfully implemented.

Sarah Box, Senior Counsellor at the OECD Directorate for Science, Technology and Innovation, is gratefully acknowledged for her comments and suggestions on the paper.

1. Introduction

In the transition to a digital era, education in G20 countries faces two challenges: reaping the benefits of AI and related technological advances to improve educational processes in the classroom and at the system level; preparing students for new skillsets for increasingly automated economies and societies, including, for some of them, the skills to contribute to the further development of digitalisation.

Digital technologies such as artificial intelligence (AI), the Internet of Things (IoT) and other advances in information and computer technology (ICT) provide opportunities to improve the education process. The education technology industry, often simply referred to as 'EdTech', is growing, with massive investments in countries such as China, the United States and India. It develops a wide range of digital solutions for education institutions and stakeholders, from online platforms to robots and smart devices. The use of digital technologies increases both the production and value of data, creating new opportunities to improve education and education policies, but also new challenges.

Education systems have started to change their curriculum and skills requirements and put a stronger emphasis on skills for innovation and citizenship in a digital era. The skills required to enter and progress in the labour market are undergoing profound changes, with more demand and emphasis on complex skills (OECD, 2019[1]).

To support the G20 AI Dialogue on "Trustworthy AI in education", this paper briefly presents promising uses of AI in classrooms and in the education system, and some possible ways to strengthen the acquisition of more complex skills such as creativity, critical thinking, communication or collaboration. It then looks at the opportunities and challenges that AI may create for educators and policy makers, aiming to stimulate debate on how countries can harness AI in their education sectors in a trustworthy way and provide sector-level insights into the implementation of the G20 AI Principles.

2. Artificial intelligence in education

Digitalisation has been one of the main drivers of innovation in education practices in the classroom in the past decade (Vincent-Lancrin et al., 2019[2]). While most innovation in the past decade related to an increased use of computers and the internet in the classroom, the next wave will be based on AI, or on combinations of AI and other technologies. A short description of AI is given in Box 2.1.

Box 2.1. Artificial intelligence

In 2019 the AI Group of Experts at the OECD defined the AI system as a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy. AI system lifecycle phases consist of: 1) planning and design, data collection and processing, and model building and interpretation; 2) verification and validation; 3) deployment; and 4) operation and monitoring (OECD, 2019[3]).

One of the most promising AI techniques is machine learning (ML), which is described as a set of techniques to allow machines to learn in an automated manner through patterns and inferences rather than through explicit instructions from a human. Behind ML is a technique referred to as 'neural networks', which is accompanied by growing computational power and the availability of massive datasets, also known as big data. (OECD, 2019[3]) In education for example, language learning applications rely on ML.

In education, artificial intelligence is embedded in many technological innovations that provide learning analytics, recommendations and diagnosis tools in various ways and for various purposes. In many cases, AI applications are still nascent and used in experimental and local contexts rather than at scale at the system level. There are, however, many examples of promising uses that foreshadow how AI might transform education in the next decades, both in the classroom and at the system levels, and address different stakeholders: students, teachers, administrators, parents, as well as policy makers. AI may particularly help achieve some of the global educational targets identified by the international community in SDG 4: "Ensure inclusive and equitable quality education and promote life-long learning opportunities for all". In line especially with the first G20 AI Principle, this section illustrates how, in the education sector, AI could be used "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."

2.1. AI applications for instruction

2.1.1. Personalising learning with AI

In terms of instruction, AI's biggest promise lies in the personalisation of learning and learning materials. Personalised learning is an educational approach aimed at customising learning based on students' individual needs and strengths. AI applications can identify pedagogical materials and approaches adapted to the level of individual students, and make predictions, recommendations and decisions about the next steps of the learning process based on data from individual students. AI systems assist learners to master the subject at their own pace and provide teachers with suggestions on how to help them. This is as important to support learning in G20 countries as in the global context. Interventions such as "Teach at the right level" are addressing this issue without technology, but the possibility to personalise learning content thanks to AI opens new initiatives to tackle this problem and improve learning outcomes in literacy and numeracy.

While those AI solutions are still relatively rare, they have been used for years by some educational establishments. In the United States, New Classrooms has for example developed the *Teach to One* math programme to personalise learning and instruction

through an intensive use of data. Implementation began in 2012 in eight schools in Chicago, New York City and Washington DC focusing on lower-secondary school mathematics. Aiming at providing instruction that is continually responsive to learners' abilities, the programme assesses students' skill levels on a daily basis and uses algorithms to target content delivery and assign students to varying instructional modes. These include teacher-led instruction, student collaborative work and educational software such as virtual adaptive tutoring. The Teach to One model relies on data from continuous formative assessment to identify individual learning gaps in maps describing progression in skills, for instance the understanding of mathematical relationships between ratios and rational numbers. Every day, students access computer dashboards displaying their progress and presenting them with tasks to work on their skills gaps as well as links to a variety of educational materials. Since skill maps are non-linear, students are allowed to move at their own pace and design their own "playlists" of tasks and skills. The large amount of data generated in this process is fed back to an underlying information system. The iteration of the model informs a daily reconfiguration of personalised learning paths and the design of broader, two-week instruction cycles. The process also provides teachers with real-time information about class and student performance through dynamic dashboards, allowing them to target their support to students' learning in a timely manner.

Other applications of AI are based on the same principle, be it in school or out of school. One of China's largest companies in out-of-school tutoring education is TAL Education Group (Beijing Century Good Future Education Technology Co., Ltd.) – a company also offering in-school services. Its AI lab has developed several types of solutions to help Chinese students prepare for their university entrance exams. The "Adaptive Test and Learning Plan" data mines its large set of assessment questions and provides students with real-time personal questions allowing to better understand their current knowledge level, and deciding which offline class suits them best. The system also uses those results to design customised study plans and push relevant materials to parents.

In India, a recent success in education technology is BYJU-The Learning App. In 2015, it released a math and science tutor for students in grades 6 to 12, then followed with one for grades 4 and 5 that also combines instruction with diagnosis and personalisation. In addition to providing video lessons, the application gauges whether the student has understood the concepts, and then takes the student either to the next level or back to basics. This is a typical example of personalised adaptive models for learning. BYJU has become the largest "unicorn" in EdTech in recent years, and reached a USD 4 billion valuation in 2018 (with 1.3 million paying users).

2.1.2. Supporting students with special needs with AI

Providing all students with a more inclusive access to education has been a persisting challenge for most countries, even more so in less affluent countries. Inclusive education is one of the global objectives promoted in SDG Goal 4, with the explicit objective to ensure equal access to all levels of education for everyone, including persons with disabilities. AI systems have already shown their effectiveness to help students with disabilities, e.g. visual or hearing impairments or impairments in social skills (language and communication), to benefit from education. For example, wearables using AI can help visually impaired students to read books and recognise faces, and thus to learn and socialise within their communities. Specialised systems have been designed to assist students with all kind of disabilities. Powered by AI, technologies such as augmented and virtual reality (AR/VR) and robotics support the learning and engagement of students with health impairments and mental health issues. While some technologies help to bypass some of the obstacles, like text-to-speech or speech-to-text applications, others are based on research and show

promising results. For example, students with autism can explore and improve social skills through interacting and collaborating with virtual characters and digital objects in a classroom.

Since 2016, students at Beijing Union University have been provided with an intelligent speech recognition system that simultaneously converts teacher's spoken language into text subtitles on a large screen. In the classroom, students with disabilities can follow the teaching through a multi-channel and multi-dimensional information input combining sign language, voice port, spoken language subtitle and text handout.

In many countries, diagnosis tools to detect special needs such as dyslexia, dyscalculia, spelling difficulties or Attention Deficit Hyperactivity Disorder (ADHD) are now based on technological devices using AI techniques (Drigas and Ioannidou, 2013[4]).

Those applications are just a few examples of how AI could help make education more inclusive and accessible in a multiplicity of contexts, regardless of peoples' disabilities or vulnerabilities.

2.1.3. Some other applications

Other applications of AI using its ability to detect patterns to provide students, teachers or parents with individualised suggestions have been developed for:

- Online and blended learning: chatbots powered by AI agents provide students and teachers with analytics on their learning.
- Classroom dynamics: different types of sensors and cameras analyse the classroom dynamics and student engagement to provide teachers with real time or post hoc feedback and suggestions.
- Foreign language learning: AI features such as speech recognition and analysis, pronunciation correction, help supplement teachers in the teaching of foreign languages.

All those application hold promises to improve the quality of education globally and improve the support and feedback offered to teachers, students and lifelong learners. They can be used in a multiplicity of contexts and by a variety of learners.

2.2. AI applications for system and school management

The algorithmic power of AI is also used to create predictive and diagnosis models to support decisions and generate feedback at the establishment (school, university, etc.) or education system level (district, region, country, etc.). AI is for example leading the way to increase completion of quality primary and secondary education, or transform tools such as standardised assessments, allowing to broaden the scope of the skills that can be assessed and increase the relevance of assessments to the skills that will become more important in a world shaped by digital technologies.

2.2.1. Reducing dropout with the help of AI

School dropout is a major educational policy issue across the globe – although countries with different levels of affluence may focus on different ages of dropout. In low income countries, 60% of upper secondary school age children were out of school in 2015. In 2018, completion rates for primary, lower secondary and upper secondary education were 68, 44 and 21%, respectively, far from the objective of universal completion by 2030. Educators and policy makers are concerned with finding the right indicators to predict dropout and

the right interventions to prevent dropout. AI systems hold promise to improve early warning systems, which are increasingly based on longitudinal datasets that are emerging in education. Even though identifying risks does not imply solving them, AI solutions help school principals to use existing data in new ways and design interventions to predict and prevent dropout more efficiently.

AI solutions are for example widespread in the United States, with a host of vendors providing districts and states with solutions helping school principals and district leaders to prevent dropout in real time. One of the virtues of these solutions is to provide feedback early enough to prevent dropout. Early warning systems typically take the form of dashboards that help visualise different types of students at risk of dropping out and (hopefully) providing them with appropriate interventions.

In less affluent countries, such as India, dropping out is also a problem and experimentation with early warning systems and interventions have also been developed and evaluated. Some of them are based on technology although the underlying data can be quite different from those used in the United States.[1]

Early warning systems are not yet a very mature technology, in spite of the promises they hold. Research on the relevant indicators to predict dropout is ongoing, as is research about possible interventions. While they hold tremendous promise, they embody the current limits of AI and the imperative to ensure it provides trustworthy and useful advice. Here the drawbacks come less from hurtful behaviours rather than from the missed opportunities presented by AI systems that do not identify accurately students who could have been helped.

2.2.2. Assessing new skills sets thanks to AI

Standardised assessments are a key feature of many education systems that can generate a lot of anxiety and drive teaching and learning practices within education systems. Increasingly, employers and policy makers feel that assessment should go beyond knowledge content and reasoning to include complex skills such as complex problem solving, collaboration, and social and emotional skills – which underpin the transformation of the world of work and of economies and societies that is recognised by the G20 AI Principles.

AI is opening new avenues in this direction. Embedded assessment, for example, creates the possibility for moment-to-moment assessments. In digital learning environments, AI systems can determine if students have mastered a specific subject. Because they store information about the student, they are also able to give formative and elaborated feedback.

AI solutions also try to assess how people think, respond to a learning situation and adapt to the students' needs and skills. Part of it can be done through speech recognition and language analysis, as well as through behavioural patterns while engaging with the task.

Game-based assessment and simulations also offer new ways to assess complex skills. They can for example incorporate assessment items into a game environment, allowing students to show their learning achievement in a playful and engaging environment. Game-based assessment can be of great value in formative assessment, adapting to the competences of an individual student, but is also applied in summative assessment. Game-based assessment typically use augmented and virtual reality as well as the adaptive power of AI. Simulations have been effectively used in science, technology, engineering and mathematics (STEM) education and in assessments in those fields, for example in

[1] See for example the Quest Alliance school dropout prevention pilot program, www.questalliance.net

medical assessments (surgical procedures). In medical education, those assessments have become common in universities.

While promising to assess certain skills, AI-based assessments face some resistance and raise new technical difficulties when used in a high stake context. The idea that tests could be different but still reliable and fair to assess people's skills challenges many students', parents' and policy makers' views about equity, showing that AI advances bring as many social and behavioural challenges to society as technical ones.

3. Skills for the digital era

The rapid adoption and diffusion of AI in the economy raises new challenges for governments and education stakeholders: what knowledge and skills should formal education systems develop given the ongoing developments? Recent research estimates that 14% of existing jobs could disappear as a result of automation in the next fifteen to twenty years, and another 32% are likely to change radically. (OECD, 2019[1]) This implies that the relative demand for skills will change, and so should the supply as well.

At this point, AI seems to best humans when it comes to repetitive and predictive tasks, tasks that hinge on computational power, classifying huge amounts of data and inputs, and making decisions based on concrete rules. (Holmes, 2019[5]) People need skills for cases where they trump the performance of machines, for example for making products and results usable for humans and communicating about them, and making decisions about abstract values. (Holmes, 2019[5]) The *OECD Skills Outlook* shows that having higher cognitive skills – literacy, numeracy or problem-solving skills in technology-rich environments, or a mix of these – significantly augments the probability that people will move from using the internet for information and communication to a diversified and complex use, taking other determinants into account (OECD, 2019[6]) ((Elliott, 2017[7]).

In the digital era, complex skills that are less easy to automate become increasingly important. Creativity and critical thinking are becoming increasingly important in the labour market, and contribute to a better personal and civic life. (Vincent-Lancrin et al., 2019[8]) The speed, volume and reach of information flows on the Internet emphasise the importance of cognitive skills. Critical thinking is of particular importance to meet some of those new demands, as students must be able to read complex texts in order to distinguish between credible and untrustworthy sources, and between fact and fiction. Creativity is also key to the development of new solutions that cannot be invented by computers yet, including solutions that are enhanced by AI and the use of robots. Besides the economic argument that complex innovation skills are less easy to automate, they also contribute to human well-being and to the good functioning of democratic societies. This is also true for socio-emotional skills such as persistence, communication or collaboration.

The G20 AI Principles acknowledge the evolving nature of the skills required to cope with the transformation of economies and societies, and recommend that governments work with stakeholders to not only empower people to use AI but also equip them with the skills that will help workers to have a fair transition as AI is deployed. While it has to continue over the life cycle, this endeavour starts in education, in school, vocational education and tertiary education.

Virtually all school curricula in OECD countries officially promote the development of creativity, critical thinking and other innovation skills in school and in higher education. This is also becoming increasingly common in G20 countries, including China and India. Yet, teachers often find it unclear what they mean and entail in their daily teaching practice. The OECD worked with school networks in 11 countries to define and operationalise

creativity and critical thinking in the school context, providing policy makers and practitioners with scaffolding tools (rubrics, lesson plans) and examples of professional development plans supporting the effective development of those skills. It showed that with some support, teachers were able to successfully adjust their pedagogy and lesson plans to develop them alongside content knowledge (Vincent-Lancrin et al., 2019[8]).

Another skills challenge for countries related to the development of AI in society lies in the opening of enough specialised tertiary education programmes preparing students for the development of those solutions. In this respect, STEM education plays a crucial role. Many of those new programmes can benefit from collaboration with the business sector. Especially in tertiary education, a focus on creativity and critical thinking is also very important: the OECD currently works with higher education institutions in 15 countries to identify good practices in this area.

4. Policy challenges

The deployment of AI in education raises new challenges for educators and policy makers. Several of those challenges relate to trust and to shaping a trustworthy use of AI, in line with the key messages of the G20 AI Principles (Box 4.1). A first challenge concerns creating and maintaining trust in AI systems. Transparency, explainability and accountability of AI systems in education are important aspects of this challenge, especially given the critical role of education in people's subsequent employment and life opportunities. A second challenge consists of ensuring the use of AI systems to serve human-centred values in protecting and securing (personal) data.

Box 4.1. G20 AI Principles

In June 2019, G20 Leaders welcomed Principles for responsible stewardship of trustworthy AI (the G20 AI Principles). These included the Principles of:

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

a. 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.

b. 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.

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:

i. to foster a general understanding of AI systems;

ii. to make stakeholders aware of their interactions with AI systems, including in the workplace;

iii. to enable those affected by an AI system to understand the outcome; and

iv. 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.

Robustness, security and safety

a. 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.

b. 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.

c. 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.

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.

The G20 also took note of five Recommendations for national policies and international co-operation for trustworthy AI, including investing in AI research and development, fostering a digital ecosystem for AI, shaping an enabling environment for AI, building human capacity and preparing for labour market transformation, and international co-operation for trustworthy AI.

Source: (G20, n.d.[9]), www.meti.go.jp/press/2019/06/20190610010/20190610010-1.pdf, (accessed 25 February 2020).

4.1. Trust in AI

As highlighted by the G20 AI Principles, reaching the full potential of AI in education will require that policy makers, educators and other stakeholders trust AI systems and their social use. This is particularly true in education as AI applications could be used to make decisions having a high stake for students: it could be used to support admission decisions, or to identify the type of support, including financial support, required by different types of learners, for instance.

Trust in AI has multiple dimensions. In education, AI might be considered trustworthy when it does properly what it is supposed to do, but also when one can trust that human beings will use it in a fair and appropriate way. For example, early warning systems powered by AI will typically profile students and identify who is at risk of dropping out. If their effectiveness in identifying the right students is too limited, even if they do no more harm than the lack of a system, they are not fully trustworthy and need improvement through further research and development. Another possibility is that they are accurate but misused. Identifying who is at risk of dropping out matters only if a good (human) intervention to support the students and address that risk is implemented. Some interventions might aim to improve completion and contribute to social justice, fairness and non-discrimination, in line with the G20 Principles. However, other interventions might seek to exclude "at risk" students from school, for example because in some accountability regimes they may lead to sanctions against the school, or a loss of reputation. It is thus not just AI that needs to be trustworthy, but also the interaction between humans and AI.

In other cases, which are still rare in most education systems, AI could lead to automatic decisions, or suggestions that will likely become decisions. For example, this could be the case for admissions to schools or universities based on certain algorithms. In some cases, this may increase fairness (for example, if the system was previously biased), but it may also have unintended consequences. As the new system will likely change the beneficiaries of the most in-demand schools, trust can only come with transparency and explainability of the criteria and algorithms. Expanding "openness" to algorithms is one solution for transparency, but for some AI techniques (such as deep learning), explainability remains difficult. Some countries (such as France) have renounced the use of some types of algorithms in public decision-making because of the difficulty in explaining them to a lay audience.

Countries may deal with questions of trust in different ways. Promoting social interaction and mutual trust is one of the pillars in China's AI strategy.[2] According to the EU guidelines for trustworthy AI,[3] AI should be transparent, that is traceable (documentation and identification of decision-making) and explainable (of both technical processes and related human decisions). Humans should have a right to be informed that they are interacting with an AI system, and capabilities and limitations should be communicated to AI practitioners or end-users. The United States' strategy on AI also confirms that further research on AI should lead to more secure, robust, and safe AI systems that are reliable and trustworthy.[4] All these strategies have aspects that closely mirror the G20 AI Principles.

4.2. Dealing with privacy and security

Societies and individuals can benefit from AI innovations. They can have a positive impact on education and learning outcomes, and prepare students for a digital future. But two other important policy issues relate to data protection and security, given that many students are minors, and to possible biases embedded in AI algorithms or in the data that feed them.

Massive data collection typically lead to concerns about privacy and data security. While the use of personal data enhances the effectiveness of AI systems in education, the collection and storage of data create new risks for privacy of students. Beyond the "Big Brother" fears that are common to all sectors of society, additional concerns related to privacy and AI in education usually are at least twofold. Families are concerned that education institutions or even employers may use "old" data to make decisions, which raises the question of how long and which data could be stored and retrieved to make some decisions. A second question relates to the possible use of the data for commercial purposes in a sphere where commercial interests are often excluded.

In many ways, those concerns relate to the "human-centred" nature of trustworthy AI, and the extent to which in a world where AI becomes more prominent, some habits and rights of humans can be preserved – for example the ability not to be constantly observed, not to risk to have one's private information publicly exposed, the ability not to be judged based on past or irrelevant (but now available) information, etc.

Countries have specific ways of dealing with privacy and security issues, which typically apply to education. In the European Union, the General Data Protection Regulation (GDPR) sets a relatively strict framework for the use of personal data. The use of data – including sharing and storage – is allowed only according to specific criteria. Transparency, data and storage limitation, and accountability are amongst the most important principles of the GDPR. In the United States, the US Family Educational and Privacy Rights Act sets out a specific framework for the use of personal data in the case of education. The Chinese Governance Principles for Responsible AI state that AI development should respect and protect personal privacy and fully protect the individual's right to know and right to choose.

[2] 'Full translation: China's 'New Generation Artificial Intelligence Development Plan' (2017)', https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/ (visited 2020, Jan. 17).

[3] 'Ethics Guidelines for Trustworthy AI', https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top (visited 2020, Jan. 17).

[4] 'Artifical Intelligence for the American People', https://www.whitehouse.gov/ai/ai-american-values/ (visited 2020, Jan. 17).

AI systems should also continuously improve transparency, explainability, reliability, and controllability, and close attention should be paid to the safety and security of AI systems.[5]

5. Conclusions

AI holds a promise to improve educational processes and outcomes in the classroom and at the system level and help achieve SDG 4. Currently, the deployment of AI in education remains limited at the system level. It is still mainly embodied in applications and solutions for individuals rather than for schools or governments. However, the EdTech industry keeps growing, and massive investments are being made in G20 countries. There is no doubt that AI will become pervasive in education and that education policy makers and governments will face the challenges of operationalising the G20 AI Principles in the education sector.

In addition to implementing the Principles for responsible stewardship of trustworthy AI in education, several policy considerations outlined in the G20 AI Principles appear of particular relevance; notably, the role of long-term public investment, and encouragement of private investment in R&D, to spur innovation in trustworthy AI, and creation of a policy environment that supports an agile transition from R&D to deployment. There is still limited evidence about the effectiveness of many AI solutions in education, a sector where both public and private investment in research and development is typically low. We do not know whether many of the current promising AI applications are technically trustworthy, nor whether they may possibly lead to non-trustworthy behaviours by some actors in the education sector. This is particularly important in a public sector where the costs of buying technology should be matched by tangible benefits. In addition, education systems rarely have establishments that allow for experimentation and the smooth transition from the research and development to the implementation phase. But new forms of innovative school networks could be developed as the traditional ways of assessing the effectiveness of educational interventions (e.g. randomised control trials) may be too slow in a rapidly evolving technological context.

Another important policy consideration is how governments can work with stakeholders to shape AI in education to help prepare for the transformation of the world of work and society. As shown above, in a different domain, it is possible to better prepare students and learners for the transformation of work and society, notably by developing their complex thinking skills such as creativity or critical thinking. Some initiatives have shown that it is possible to work with stakeholders to achieve those goals, but also that change will not happen by itself.

Finally, a particularly challenging aspect of the G20 AI Principles lies in considering the "open sharing" of data with companies, which is particularly difficult in the education sector – although it happens increasingly in a controlled way in many countries. New solutions will have to be invented, perhaps inspired from developments in other public sectors.

[5] 'Full translation: China's 'New Generation Artificial Intelligence Development Plan' (2017)', https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/ (visited 2020, Jan. 17).

References

Drigas, A. and R. Ioannidou (2013), "A Review on Artificial Intelligence in Special Education", in *Information Systems, E-learning, and Knowledge Management Research, Communications in Computer and Information Science*, Springer Berlin Heidelberg, Berlin, Heidelberg, http://dx.doi.org/10.1007/978-3-642-35879-1_46. [4]

Elliott, S. (2017), *Computers and the Future of Skill Demand*, Educational Research and Innovation, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264284395-en. [7]

G20 (n.d.), *G20 Ministerial Statement on Trade and Digital Economy*, https://www.meti.go.jp/press/2019/06/20190610010/20190610010-1.pdf. [9]

Holmes, W. (2019), *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*, Center for Curriculum Redesign. [5]

OECD (2019), *Artificial Intelligence in Society*, OECD Publishing, Paris, https://dx.doi.org/10.1787/eedfee77-en. [3]

OECD (2019), *OECD Employment Outlook 2019: The Future of Work*, OECD Publishing, Paris, https://dx.doi.org/10.1787/9ee00155-en. [1]

OECD (2019), *OECD Skills Outlook 2019 : Thriving in a Digital World*, OECD Publishing, Paris, https://dx.doi.org/10.1787/df80bc12-en. [6]

Vincent-Lancrin, S., C. González-Sancho, M. Bouckaert, F. de Luca, M. Fernández-Barrerra, G. Jacotin, J. Urgel and Q. Vidal (2019), *Fostering Students' Creativity and Critical Thinking: What it Means in School*, Educational Research and Innovation, OECD Publishing, Paris, https://dx.doi.org/10.1787/62212c37-en. [8]

Vincent-Lancrin, S., J. Urgel, S. Kar, and G. Jacotin (2019), *Measuring Innovation in Education 2019: What Has Changed in the Classroom?*, Educational Research and Innovation, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264311671-en. [2]