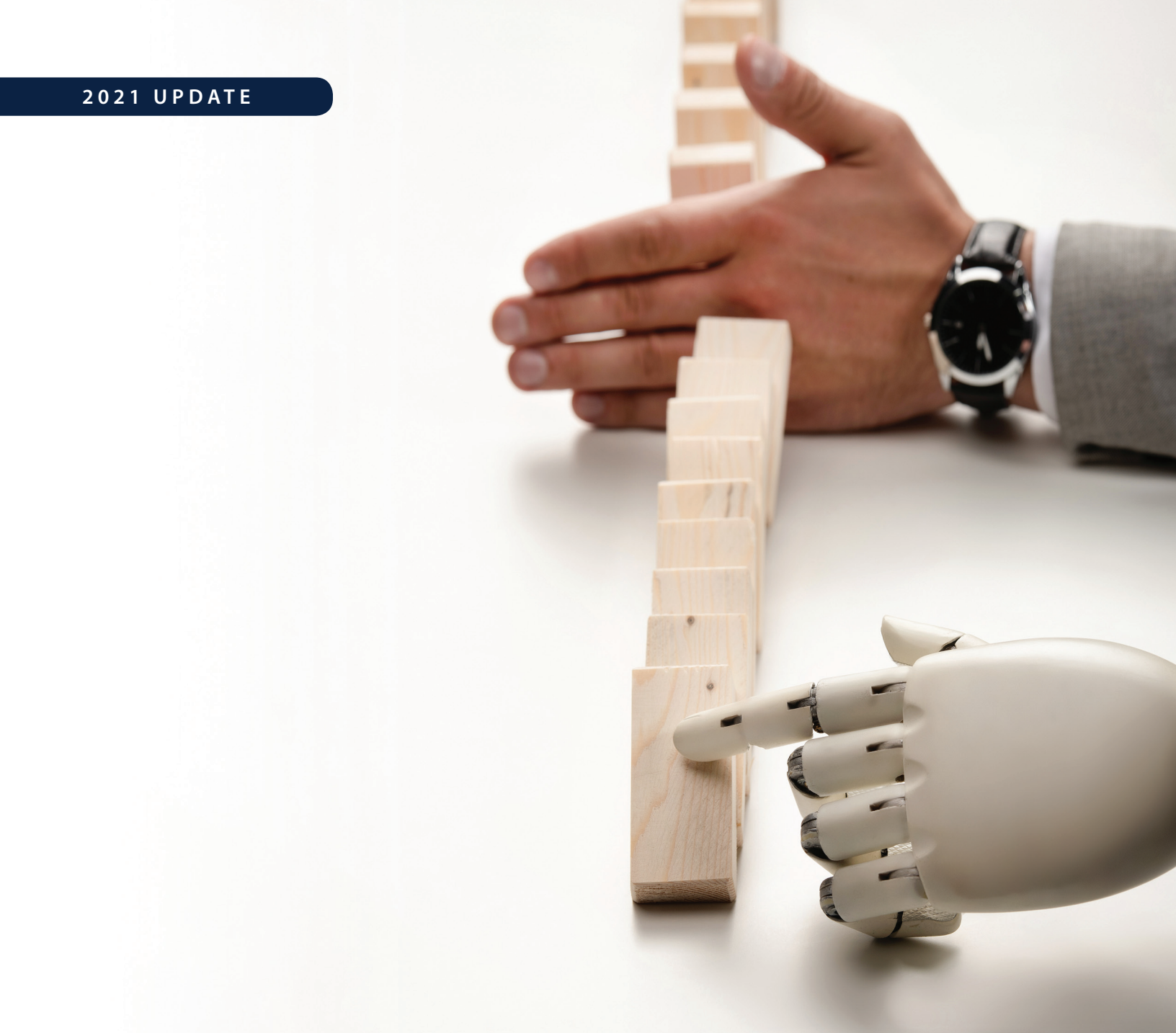
2021 UPDATE



Responsible AI

A GLOBAL POLICY FRAMEWORK

ETHICAL PURPOSE *SOCIETAL BENEFIT* *ACCOUNTABILITY* *TRANSPARENCY* *EXPLAINABILITY* *FAIRNESS* *NON-DISCRIMINATION*
SAFETY *RELIABILITY* *OPEN DATA* *FAIR COMPETITION* *PRIVACY* *INTELLECTUAL PROPERTY*

ITECHLAW®
INTERNATIONAL TECHNOLOGY LAW ASSOCIATION

Responsible AI

A GLOBAL POLICY FRAMEWORK

2021 UPDATE

John Buyers and Susan Barty, Editors

ITECHLAW®
INTERNATIONAL TECHNOLOGY LAW ASSOCIATION

McLean, Virginia, USA

This book does not provide legal advice. It is provided for informational purposes only.

In the context of this book, significant efforts have been made to provide a range of views and opinions regarding the various topics discussed herein. The views and opinions in this book do not necessarily reflect the views and opinions of the individual authors. Moreover, each of the contributors to this book has participated in its drafting on a personal basis. Accordingly the views expressed in this book do not reflect the views of any of the law firms or other entities with which they may be affiliated. Firm names and logos, while used with permission, do not necessarily imply endorsement of any of the specific views and opinions set out herein.

The authors have worked diligently to ensure that all information in this book is accurate as of the time of publication. The publisher will gladly receive information that will help, in subsequent editions, to rectify any inadvertent errors or omissions.

International Technology Law Association
7918 Jones Branch Drive, Suite 300
McLean, Virginia 22102, United States
Phone: (+1) 703-506-2895
Fax: (+1) 703-506-3266
Email: memberservices@itechlaw.org
itechlaw.org

Cover and chapter title page designs by Stan Knight, MCI USA
Text design by Troy Scott Parker, Cimarron Design

This book is available at www.itechlaw.org.

ISBN-13: 978-1-7339931-1-1

Printed in the United States of America.

Printing, last digit: 10 9 8 7 6 5 4 3 2 1

Contents

Preface

In May 2019, almost a year to the day after we had commenced our initial collective efforts, ITechLaw published the first edition of *Responsible AI: A Global Policy Framework.* As editor and contributor to the first edition, I had the great honour to work with a remarkable multi-disciplinary team of 54 technology legal experts, researchers and industry representatives from 16 countries to produce a richly researched policy guide to the responsible deployment of AI systems.

As noted in the first edition of *Responsible AI,* the policy framework that we published in 2019 was necessarily embryonic. Artificial intelligence's development is still in its infancy and the potential societal impact of artificial intelligence is difficult to fully grasp, particularly in a field in which the rate of change continues to be almost exponential. These factors have placed a great weight of responsibility on all those who are engaged in the development and deployment of such AI systems. It is not surprising, therefore, that not only policy makers, but also industry representatives and AI researchers are looking for solid legal and ethical guideposts. We are, collectively, participating in an ongoing dialogue.

It is in this context that I am pleased to welcome the publication of the 2021 Update to *Responsible AI: A Global Policy Framework.* As we undertook to carry on the dialogue, we could not have been better served than by the two editors of this current update, John Buyers of Osborne Clarke LLP, UK and Susan Barty of CMS of CMS LLP. Together with a team of 38 specialists from 17 countries, John and Susan have not only produced a substantive update to each of the eight principal chapters to *Responsible AI* and a comprehensive update to the original Global Policy Framework, but have also developed a practical "Responsible AI Impact Assessment" template that we hope will be of significant value to AI experts and industry leaders.

This Update continues to fulfill the promise and potential of ITechLaw as a global association promoting networking and thought-leadership amongst leading technology lawyers worldwide.

– Charles Morgan
McCarthy Tétrault LLP
President, International Technology Law Association
February 2021

Foreword

It would not be an understatement to say that the world has changed beyond recognition since the publication of the first edition of *Responsible AI.* We have all been placed in the grip of a global pandemic, dramatically changing our working and personal lives, forcing distance between us and our loved ones and transforming innocent gestures of social interaction, such as shaking hands and hugging, into potentially deadly interactions. Where once we might have flown or driven to a meeting or conference, we now use video conferencing.

Isolation has made us even more dependent upon technology: to work, to socially interact, to inform, educate and to entertain. Social media and predictive technologies have become ever present in ways we could not even have imagined: driving and manipulating opinions, influencing behaviours and inevitably powering news cycles. Indeed, as we bring this update to publication we're witnessing at first hand the impact of these technologies on a very unconventional US Presidential election.

The consensus is that rather than enrich us as human beings, exposure to too much technology diminishes us. This is perhaps not surprising as forced isolation has driven many to the conclusion that we need real social relationships and interaction to thrive as human beings.

It is in this environment that we bring you our 2021 update to *Responsible AI.* In a fast moving world, Artificial Intelligence moves at light speed. We're now seeing the first nascent global steps towards regulation: the collective governmental realisation of the enormous harm that this technology can wield if left untrammelled. It looks like the EU is "first out of the blocks" with a proposal that would align machine learning to a regulatory environment not too dissimilar to the one Europeans face with data. The EU's compliance driven thinking is inevitably tempered by the more entrepreneurial and enterprise friendly approaches advocated by the United States and China. Time will tell which vision will prevail.

In the meantime, it has become ever more critical to measure and gauge the impact of artificial intelligence "on the ground" and away from academic debate. We are inevitably "wising up" to the consequences of ill thought through development and use– whether that is physical harm, exclusion or erosion of personal liberty. It is in this environment we launch our Responsible AI Impact Assessment tool (or RAIIA for short) which is designed to help measure, in quantifiable and real terms, the impact of a proposed AI solution. We hope you find it a valuable, and practical tool.

"Responsible AI" is a unique and precious initiative of ITechLaw—it drives our collective organisation to heights that others do not reach and showcases the intellectual vision of its members. We consider ourselves to be privileged and humbled to have worked on this update with such a wise group of international friends. Alongside the pressures and inevitable strain of editing it, it has provided us with invaluable insight and companionship. We look forward to the time when we can all meet together again. In the meantime, we wish all of our readers health and success (and of course further insight into the complexities of artificial intelligence) for what is hopefully a brighter 2021.

– John Buyers, Osborne Clarke LLP
Susan Barty, CMS LLP
February 2021

Updates to the First Edition

Responsible AI

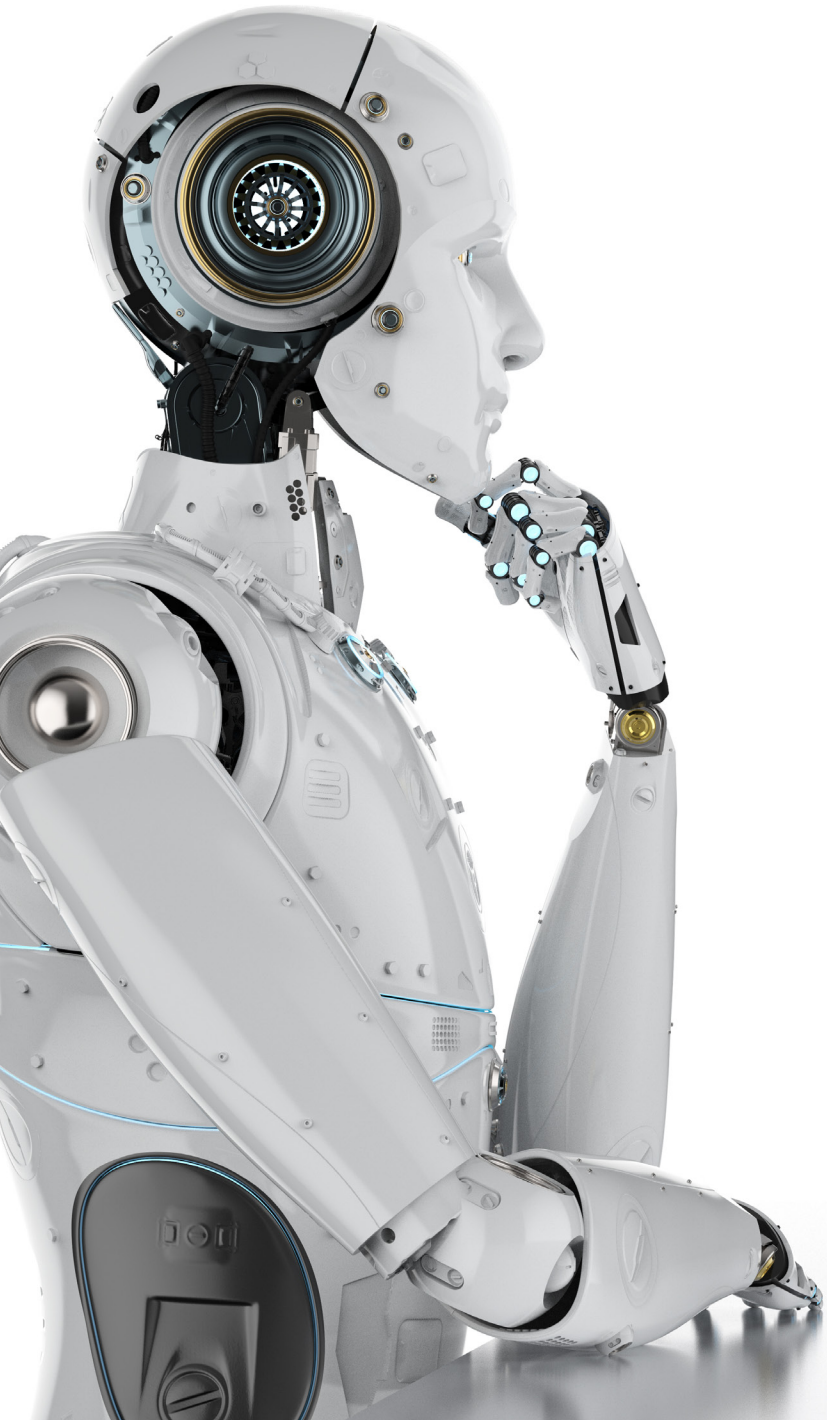
A GLOBAL POLICY FRAMEWORK

Responsible AI

A GLOBAL POLICY FRAMEWORK

Principle 1

ETHICAL PURPOSE AND SOCIETAL BENEFIT



Principle 1 Commentary

ETHICAL PURPOSE AND SOCIETAL BENEFIT

CHAPTER LEAD
Patricia Shaw | Beyond Reach Consulting Limited, UK

Francis Langlois | McCarthy Tétrault, Canada

Charles Morgan | McCarthy Tétrault, Canada

Steven De Schrijver | ASTREA bv cvba, Belgium

Alesch Staehelin | TIMES Attorneys, Switzerland

Human agency and autonomy

Ethical concerns were at the core of our reflection on artificial intelligence in the first edition of *Responsible AI.* We deemed essential that "an ethical purpose, a purpose that has a demonstrable and reasonable societal benefit" remain ever-present in the mind of jurists, legislators and policymakers working on the foundation of the law of AI.

In that context, we based our reflection on the ethical concepts of beneficence and non-maleficence, and then focused on four areas where the potentially transformative impact of AI is a matter of significant societal debate: a) the transformation of the workplace; b) the ecological impact of AI; c) the militarised uses of AI, especially in the form of lethal autonomous weapons systems; and d) the spread of AI-powered fakes news, deep fakes and disinformation. Our objective when discussing these issues was to explore concrete examples of ethical issues that arise in the context of the development of and deployment of AI systems and to insist upon the importance of giving due consideration of such issues prior to deployment.

In this update to Principle 1 of the Responsible AI framework, we provide context for the inclusion of a new subsection 2 for this first principle, expanding the reflection commenced in the first edition of this chapter with a specific focus on the core themes of human autonomy and human agency, which implicitly underlay several of the examples of ethical tension previously discussed.[1] How do AI systems affect us directly as humans? Moreover, to what extent should we allow AI systems to transform our current human condition and our social world? What are the risks that humans will be inappropriately controlled by technology in a manner that *threatens* our autonomy and agency instead of serving as a valuable tool that *enhances* them? How can we mitigate against such risks?

Below, we explore these questions through the lens of two basic questions:

- **AI-powered surveillance:** When does protective oversight or efficiency-enhancing attentiveness become dangerous surveillance?

- **AI-driven behavioural control:** When does a helpful AI-enhanced suggestion become inappropriate manipulation?

Of course, these are not easy questions to answer and different people and different cultures may answer them differently. Nevertheless, we would argue that there is, in each case, a line that should not be crossed and hence that, prior to developing, making available or using an AI system, the fundamental questions should be posed: "Will this AI-system enhance or threaten human autonomy and human agency?" and "How does this impact on human dignity?" at home, in public and in the workplace.

AI-Powered surveillance: Self-censorship and loss of independent thinking and expression

In 1890, Samuel Warren and Louis Brandeis described in their famous article on the Right to Privacy, which developed the "right to be let alone," that: *"numerous mechanical devices threaten to make good the prediction that 'what is whispered in the closet shall be proclaimed from the house-tops.'"*[2] Pre-dating fundamental rights, US law recognised that privacy needs bespoke protection in the face of invasive technology.[3] In attempting to meet the problems posed by the technological and social changes occurring in their days, the US courts progressively devised a tort of invasion of privacy[4] and the right to be let alone (for which no parallel tort seemingly existed under UK law). Subsequently, the US lawmakers enacted the 1965 Restatement of Torts (2nd)[5] which recognised the tort of "Intrusion upon the plaintiff's seclusion or solitude, or into his private affairs" amongst other privacy centric torts.

But if Warren and Brandeis decried the intrusive nature of "modern" technology upon our private sanctuary in 1890 (the (then) recent invention of photography and its use by a sensationalist press), what would they think of technology's ubiquitous intrusions today! Indeed, in modern times nearly every commercial street and building have CCTV cameras permanently watching our every movement. An average American is caught on CCTV camera an estimated 75 times a day, while the average Londoner holds the record of being photographed and filmed 300 times a day.[6]

Warren's and Brandeis's alarming description of intrusive "mechanical devices" is even more relevant in relation to the surveillance exercised by "always on" technology that we increasingly bring into our homes and close to our bodies, such as virtual assistants, smart home connected devices, wearables and, most frequently, smartphones. The information yield of such technologies is exponentially increased when combined with big data and AI. While the analysis of all this information would be daunting for human beings, one of the most significant uses of artificial intelligence is in the mining of vast databases to extract precious insights, notably on human behaviour. All these technologies allow for greater intrusion than peaking over a fence with a camera; by virtue of being in our pockets or in our living rooms—and almost permanently connected to the Internet—they give access to increasingly intimate aspects of our lives. As Yuval Noah Harari argues, we have moved from "over-the-fence" surveillance to "under-the skin" surveillance.[7]

The benefits of surveillance

Use of such algorithmic systems can provide real societal benefits, notably in the form of actionable predictions. In the private sector, this may result in tailored content, concentrated pools of information and more accurate search results. Consumers can be shown only products that are appropriate and suitable to their specific needs and tastes (a movie to watch on Netflix, for instance),[8] and offered services (such as credit cards, loans and insurance) for which they would be eligible. Other beneficial use of AI and big data include FaceID that conveniently unlocks a user's smartphone based on its machine learning algorithms which compare an instant scan of the user's face with the scan that is stored. Virtual assistants can help to get directions while driving or may draft text and email messages. Smart thermostats can adjust the temperature in houses automatically. In a society where time is of the essence, these AI tools facilitate many daily tasks, making them less time-consuming. Moreover, as we have seen more recently, AI-enhanced technologies may play an essential role in helping society respond efficiently to the COVID-19 public health and economic crisis, notably through the use of machine learning-based contact tracing apps.[9]

In the public sector, automated decision-making has grown to power decisions that impact lives and societies.[10] With algorithmic systems, governments can ensure appropriate and relevant notifications, advice and services are delivered as effectively as possible to citizens. They create efficiencies, save time (and money), and make access to information and products/services more convenient. Additionally, the use of technologies such as CCTV or license plate readers by public authorities, especially for surveillance purposes, is in most cases based on legitimate reasons of societal benefit such as prevention and control of criminal offences, security or safety requirements or public health. Smart cities may also use AI surveillance to improve traffic flow by, e.g. changing traffic light phasing in response to real-time activity.[11] Recent studies show that already 75 out of 176 countries globally are using AI technologies for surveillance purposes.[12] As another example, in reaction to the COVID-19 pandemic, several governments, with support from the private sector, are venturing to augment contact tracing with AI capacities in the hope to more efficiently control the spread of the virus.[13]

The downsides of AI-driven surveillance systems

The development of such technologies can also lead to losses in privacy and autonomy as well as to infringement upon fundamental rights. As a result of the shocking revelations of Edward Snowden, for example, we learned that the NSA could monitor essentially every telecommunication in the world. Imagine the consequences if such surveillance powers were extended beyond the traditional Internet or telephone communications to the billions of IoT devices with which we interact, consciously and unconsciously, at all times. In addition, the combination of contact tracing and AI, notably through the use of smartphones applications taking advantage of location data, has been met with concerns over increased surveillance.[14] In other words, gains in efficiency or security have a high cost: the loss of sanctuary and ubiquitous surveillance.

Like Warren and Brandeis who worried about the impact of photography on the right to be let alone, there is an increasing concern that AI technology could adversely affect human behaviour. As Edward Snowden has said, the absence of privacy is not the presence of security, but it is rather the presence of censorship. China serves as a prime example of how public use of AI-driven surveillance measures may have gone too

far, even though it may be based on culturally legitimatised reasons of security and public safety. While its facial recognition system can recognise offenders that ignore a red light when crossing the street, which is said to be a large problem in China,[15] certain reports claim that AI facial recognition technology is programmed in China in a way as to recognise members of certain minorities such as Uighurs based on their appearance, which then keeps records of their comings and goings. This raises concerns on the possible racial profiling which AI can cause to happen.[16] Regarding facial recognition, there is also currently a wide societal debate in countries like the US over the use of such technologies for law enforcement purposes. The City of Boston, for instance, considered a ban on the use of facial recognition technology, notably due to the unreliability of present-day AI software when identifying people with darker skin tones.[17] Moreover, following the killing of George Floyd, companies such as Microsoft, Amazon and IBM announced they will refrain for selling facial recognition systems until proper legislation is put in place.

The use of AI for policing purposes is not limited to facial recognition. AI surveillance is also used for predictive policing, whereby algorithms analyse historical data on crime to detect where further acts are the most likely to happen. Based on this data, people with characteristics that correlate with criminal behaviour will more likely be policed, even though there is absolutely no guarantee that these persons will develop any future criminal behaviour. Although innocent, such persons will carry the burden of being additionally subject to surveillance.[18]

Moreover, one of the secondary impacts of the COVID19 crisis is displacement of the surveillance occurring in the workplace to the new de-facto office for many workers: the home. Workers who, prior to the lockdown, had had to login to the IT system at their desks with retinal scan or facial recognition technology, that worked with IT systems able to monitor the amount of time they spend at their desks and measure their productivity,[19] accompanied by an virtual 'open door' (i.e. always online and accessible) culture of internal communication are now bringing all this technology home. The move to remote working from home, has made the tacit amount of surveillance in the workplace stark. In some cases, parts of the surveillance have merely swapped location, now being willingly carried out by workers from their very homes, leaving even less of a divide between work and home. This begs the question: "how much workplace surveillance is too much?"

The impact on human autonomy

Both private and public use of AI-driven technology for surveillance purposes may pose a serious threat to human autonomy, which is an individual's capacity for self-determination or self-governance. The self-determined actions of individuals may become impacted by an outside influence, even though the individual is unaware of its existence. But even if the individual has reason to believe that such outside influence exists, it may be very difficult to prove this due to the lack of transparency of surveillance systems.[20]

In turn, the feeling of being under surveillance (whether true or not) may lead to a further disturbing impact on the individual: the growth of distrust or even the inability to trust. Individuals may adapt their behaviour as they take into account that they are being subject to surveillance, whereby such behaviour may even become the new normal. In the worst case, certain individuals may develop paranoia or other

mental health issues (e.g. anxiety may increase which can lead to high blood pressure, obesity, respiratory problems[21]).[22]

Surveillance, whether by the government or by private actors, may lead to (un)conscious self-censorship. Research into the online behaviour of US citizens following the Edward Snowden revelations on government surveillance led to a clear decline in Wikipedia searches for certain terrorism-related keywords (e.g. Al Qaeda, chemical weapon and jihad).[23] Such self-censorship also weakens one of the strengths of a healthy democracy, namely the freedom of speech which also includes voicing concerns over political and social questions.[24] But self-censorship may also affect inter-human relationships, as people that know they are being watched may also think twice about their communications with others as they may be afraid that their messages could be taken out of context. Consequently, people may be less willing to foster real intimacy and shared understandings.[25]

This underscores the importance of developing comprehensive and appropriate legal regimes in order to ensure AI systems are used in a beneficent way that protects human autonomy and agency. Organisations that develop, make available or use AI systems require guidance as to when one crosses the line between protective oversight or efficiency-enhancing attentiveness to dangerous surveillance that threatens human autonomy.

The EU's Ethics Guidelines for Trustworthy AI mention in this respect that *"humans interacting with AI systems must be able to keep full and effective self-determination over themselves, and be able to partake in the democratic process."* Instead of coercing or deceiving humans, it is important that AI systems are designed in a way which augments, complements and empowers human cognitive, social and cultural skills.[26] In a time where data on a person's life is more easily available than ever, policymakers must make sure that the wide possibilities to gather such data and to subject people to surveillance by the devices they use or which the authorities may use in public spaces are bound by strong legal and ethical frameworks.

Beyond policy interventions, technologists can develop new applications that consider the preservation of human autonomy and agency from the design stage. For example, in the midst of the COVID-19 outbreak, the Montreal Institute of Learning Algorithms (MILA) proposed a contact tracing app called COVI Canada App. Although the project ultimately did not come to fruition, its design approach was remarkable for the various ways by which MILA sought to preserve user privacy and human agency. Its multi-layered approach combined cryptographic messaging for the transfer of data, as well as on-device storage and daily deletion of most of the data. Moreover, it included pseudonymization of personal data and the creation of a data trust to ensure independent governance. Finally, rather than assuming consent, the COVI App proposed a "multi-layered, progressive disclosure approach" which would have used methods like graphics and illustrations to make clear the privacy implications of its system. The COVI App was thus a good example of how agency-enhancing mechanisms can be combined with privacy measures to create more trustworthy AI systems.[27]

In this context, we have introduced Sections 2.1 and 2.2 to principle 1 of the Responsible AI framework in an effort to require organisations that develop, make available or use AI systems that surveil human behaviour to implement safeguards:

- to promote the right to be let alone, informed human agency and autonomy;
- to avoid destructive self-censorship, loss of individuality and identity, loss of freedom of expression;
- to provide full transparency as to whether and when a device's voice, movement or image surveillance features have been activated; and
- to store sensitive personal data collected locally by IoT devices (such as fitness monitors and smart phones) and natural language, movement and image data collected by "always on" IoT devices (such as personal assistants and smart home devices), to the greatest extent possible, in encrypted format, only locally on the device in a manner that allows for the maximal level of autonomy and control over the data by the individual(s) to whom it relates.

AI-driven behavioural control: From empowerment to manipulation

Human autonomy, and freedom of choice (brain hacking and attention deficit)

Surveillance is not the only way AI and big data can impact human autonomy. The flow of information can also be reversed: once new insights are gained about consumers and citizens, corporations and governments can use this information to influence behaviour. This, of course, as always been the goal of advertisement or propaganda. As we will see, however, the use of data driven algorithmic systems to generate incremental timely messaging and targeted advertising has led to an interference with self-determination. By using behavioural data, predictive analytics and inferred data, organisations have been able to nudge decision making. The timing of that messaging can be predicated, for maximum impact, on an individual's browsing/viewing habits or other triggers, such as household or car insurance renewal dates. Such timely reminders can act as useful prompts to engage with our service providers. However, whilst such messaging can be useful, it can adversely impact human autonomy (use of memory recall, critical thinking and through inciting thoughts and feelings and depriving individuals of attention).

Today, algorithmic systems, like the ones discussed above, are used to monitor, track, assess, categorise and analyse online behavioural and in-app activity data. This data is referred to by Shoshana Zuboff as our "Behavioural Surplus."[28] It tell companies and governments information that we do not even know about ourselves, information that is powerful and can be used, either for us or against us, with or without our knowledge or awareness of it, to modify our online, in-app or offline behaviour.

The patterns recognised from this behavioural surplus are being used to predict with high levels of accuracy an individual's next move: what they will buy, watch, read and what and where they will exercise, and how they will vote,[29] amongst other attributes. Once known, predictive insight can be used in probabilistic modelling which in turn can give greater certainty to predictions about our future activities, producing "economies of action" and a "behavioural futures market."[30] This shows where prediction analytics can be autonomy-invasive by affecting an individual's or even a group's freedom of choice.[31] As Edward Snowden put it: "Once you go digging into the actual technical mechanisms by which predictability is calculated, you come to understand that its science is, in fact, anti-scientific, and fatally misnamed: ***predictability is actually manipulation***.... a mechanism of subtle coercion."[32]

The aim of recording this kind of information is ostensibly to enhance user experience. By having a greater understanding of the thoughts, words and deeds as well as future needs of individual users, they can be given a truly personalised offerings and experiences. A nudge can provide an algorithmically driven but behaviourally informed approach to help individuals, companies and policy makers save time and money. Provided informed consent has been given, nudging assisted by activity data can be done so legitimately with proper delegated human agency seeking to respect and preserve choice.[33]

However, where this activity data is recorded without the informed consent of the persons concerned or prediction analytics are applied without the user being aware or understanding the consequences of its application, the legitimacy that may have once been provided through lawful contractual consent starts to wane. Whether it be done by private enterprise or government actor, this kind of interference with our choices impoverishes an individual's private existence and commoditises human beings[34]—the data representing a digital extension of our human selves, a digital twin, a part of us as our data self.[35]

Whilst organisations may use this data to deliberately manipulate choices, they can also use algorithmic systems to create addiction[36] and dependency, whether it be on a particular game, app or social media platform. The aim is to keep users in the product for as long as possible, vying for the user's time and attention, or to keep the user coming back for more. There is an "attention market," where economic actors broker for human attention.[37] The motivation is money—generated through advertisements, click-rates, and sales—and predictability only enhances the success rates.

This phenomenon is not entirely new. In his book *The Attention Merchants*, Columbia Law School professor and *New York Times* columnist Tim Wu tells the story of the competition for our attention, from the penny press of the 19th century, to the television of the post-war era, all the way to the age of the Internet and Social Media.[38] Printed newspapers, radio shows and television programs have long been designed to appeal to certain audiences in the hope they would be receptive to ads selling certain products and services. The process, however, was crude and imprecise. This changed with the advent of Big Data, AI and the access by technology companies to the flow of information coming from our activities on the Internet and on our connected devices. This led to the highly targeted advertising most Internet users experience every day. But as AI technology matures, it will increasingly impact our offline lives as well.

Combined with augmented reality, this could lead to a future where the struggle for our attention increases and reaches new realms of our lives. Being deprived of attention in this way, weakens relationships, causes attention deficit and erodes freedom of choice for the user. This leaves our thoughts hijacked and, as a result, our consequent actions are no longer free from outside influence.[39]

While there exists a vast body of laws concerning privacy, the invasion of human autonomy and self-determination in the form of behavioural manipulation appears to slip between the gaps of human rights as well as data protection, and consumer protection laws. This creates a captive audience which was once perhaps at first exercised through voluntary choice (not necessarily informed consent), but has increasingly become involuntary and coercive, leading to what can only be thought of as an "attentional intrusion," "attention theft"[40] or "brain hacking."[41]

AI, human autonomy and the law

Fundamental rights become mere shells without the ability by individuals to exercise meaningful freedom of choice (such as Article 9 of the European Convention on Human Rights which provides for an unqualified right to freedom of thought, conscience, and religion). Accordingly, "conduct which reduces this freedom of choice—whether improper pressure, taking advantage of individuals with a reduced capacity to choose, or the negation of individual choice implied by 'brainwashing'—constitutes a violation of that right."[42]

Nudge economics has been up until now seen as acceptable for use by regulators and businesses alike. Now with the use of AI and data driven technologies, it is unclear when digital nudge economics[43] ends and manipulation begins.

Through a lack of understanding of the underlying technology coupled with the view that Artificial Intelligence is too complex to be regulated by legislators and regulators, a lack of test cases and application of existing laws and Human Rights convention treaties to new business and governmental technological practices, these practices which interfere with freedom of choice have been left unchecked. Low levels of accountability and transparency, with undue regard for the representativeness of data or the processes put in place to address and mitigate bias, have been permitted to subsist for too long.

Civil and criminal laws currently do not explicitly address the kind of individual harms raised above where they are not intentionally deceptive or involve physical or financial harm. Individual harms arising from "seizure of attention and consequential cognitive impairments" or a bargain entered into through coercion and manipulation of thought of feelings, are intangible and therefore not addressed.

This raises questions of whether Governments should be looking to promulgate new Digital Human Rights or provide for greater Digital Consumer and Data Protections laws to identify, deter and safeguard against such violations. Whether it should be left to the Judiciary and the court system (where jurisdictions are so configured) to invoke at law a duty of care (i) to exercise good faith and non-manipulation, (ii) to not engage in algorithmic nuisance,[44] or (iii) to give effect to the autonomous individual by securing the inviolate person.[45] Alternatively, will justice be seen in equity by extending our current understanding of what constitutes an undue influence or an unconscionable bargain. Either way safeguards need to be put in place to clearly define the parameters of AI and data-driven technologies and protect against these new kinds of harm. In other words, just as Warren and Brandeis pioneered the field of privacy law in the US in reaction to the rise of the penny-press and photography, our daily interactions with the attention market, AI-driven behavioural analysis and Big data should lead to legal innovation that will ensure those technologies evolve in a way beneficial to humans.

In this context, we have introduced Section 2.3 to principle 1 of the Responsible AI framework in an effort to require organisations that develop, make available or use AI systems put in place appropriate safeguards to promote informed human agency and autonomy and to avoid destructive psychological and behavioural manipulation, addiction, dependency and attention deficit.

Agency and dignity in the workplace

Finally, before completing this update, we return briefly to one of the topics that we discussed in the first edition: the impact of AI on the workplace. A substantial and growing concern is that the quality and type of work being supplemented by AI is having an ethical impact on individuals and society.

Clearly, a vast array of technological tools, including AI-enhanced tools, have empowered individuals in the workplace by increasing their efficiency, providing remarkable new means of collaboration, as well as access to lifelong learning. Some tools we have seen come into life during the COVID-19 crisis, such as

- AI trawlers used on social media platforms to reduce disinformation and fake news, have been essential to curb inaccurate or false claims of remedies, to help save lives; and
- Chatbots have continued customer services operations, replacing traditional call centres.

The quality of work may impact on human beings through lack of challenge, dignity, fulfilment or purpose in the work. Work unable to be accurately fulfilled by AI (such as tagging, image labelling or deciphering the nuances and connotations of language) being left to humans may be repetitive and menial, but could also be harmful to the mind. The impacts resulting in PTSD, poor mental health, dissatisfaction, lack of raison d'être and/or purpose, and stifled joy and creativity.[46]

In this context, an issue that has received an increasing level of scrutiny as regards the emotional and psychological health of the workforce relates to manual content monitoring. Monotonous data labelling or image recognition of extreme and/or horrific content cause various mental problems for employees.

Low-paid content moderators are constantly facing traumatic images and videos. Studies show that many cope with that by telling dark jokes about committing suicide and by "self-medicating" with illicit drug use to "numb" the impact. Team leaders micro-manage content moderators' every bathroom break. Employees are developing PTSD-like symptoms after they leave the company, but are no longer eligible for any support from their former employers.[47] Some employees have begun to embrace the fringe viewpoints of the videos and memes that they are supposed to moderate: A group of current and former contractors who worked for years at a Berlin-based Internet content moderation centre has reported witnessing colleagues become "addicted" to graphic content and hoarding ever more extreme examples for a personal collection. They also said others were pushed towards the far right by the amount of hate speech and fake news they read every day. They describe being ground down by the volume of the work, numbed by the graphic violence, nudity and bullying they have to view for eight hours a day, working nights and weekends, for "practically minimum pay." A little-discussed aspect of such content moderation was particularly distressing to the contractors: Vetting private conversations between adults and minors that have been flagged by algorithms as likely sexual exploitation.[48]

Content moderators complain that their employers do not provide adequate support to address the psychological consequences of the work. They said that they could not confide in friends because the confidentiality agreements they signed prevent them from doing so, that it is tough to opt out of content that they see, and that daily accuracy targets create pressure not to take breaks. The tech industry has acknowledged the importance of allowing content moderators these freedoms—in 2015 signing on to a

voluntary agreement to provide such options for workers who view child exploitation content, which most workers said they were exposed to.[49]

In this context, we have introduced Section 3.4 to principle 1 of the Responsible AI framework in an effort to require organisations that develop, make available or use AI systems that surveil or influence employee behavior in the workplace shall put in place appropriate safeguards to promote the informed human agency, autonomy and dignity of employees and to avoid inappropriate or destructive impacts on the emotional or psychological health of employees, such as monotony of tasks, excessive surveillance, gaming of behavior, continuous exposure to horrific content.

In conclusion, AI systems can be powerful tools that empower individuals to make better informed and life-enhancing choices for our individual and collective benefit. They can also threaten us and cause (directly or indirectly, intentionally, or unintentionally) individual and collective harm by undermining human autonomy, agency and dignity. The ethical and societal risks of any AI system are multi-dimensional and are often not straight forward. Given the central importance of these issues to the flourishing of human society, it remains critically important that organisations ensure that they thoroughly assess the ethical implications and societal benefit of a proposed AI system as part of a structured Responsible AI Impact Assessment prior to its development, deployment or use.

Principle 1
Ethical Purpose and Societal Benefit

Organisations that develop, make available or use AI systems and any national laws or industry standards that govern such use should require the purposes of such implementation to be identified and ensure that such purposes are consistent with the overall ethical purposes of beneficence and non-maleficence, as well as the other principles of the Policy Framework for Responsible AI.

1 Overarching principles

1.1 Organisations that develop, make available or use AI systems should do so in a manner compatible with human agency, human autonomy and the respect for fundamental human rights (including freedom from discrimination).

1.2 Organisations that develop, make available or use AI systems should monitor the implementation of such AI systems and act to mitigate against consequences of such AI systems (whether intended or unintended) that are inconsistent with the ethical purposes of beneficence and non-maleficence, as well as the other principles of the Policy Framework for Responsible AI set out in this framework.

1.3 Organisations that develop, make available or use AI systems should assess the social, political and environmental implications of such development, deployment and use in the context of a structured Responsible AI Impact Assessment that assesses risk of harm and, as the case may be, proposes mitigation strategies in relation to such risks.

2 Human Agency and Autonomy

2.1 Organisations that develop, make available or use AI systems that surveil human behavior shall put in place appropriate safeguards to promote the right to be let alone (the right not to be subject to arbitrary interference with his privacy, family, home or correspondence), informed human agency and autonomy and to avoid destructive self-censorship, loss of individuality and identity, loss of freedom of expression and the loss of human ability to think freely and independently. Such safeguards shall include conducting a responsible AI ethical risk assessment of the technology as part of an accountable governance process prior to deployment of the AI System and ensuring that any such deployment is consistent with respect for other principles of the Policy Framework for Responsible AI such as Transparency and Explainability, Fairness and Non-Discrimination, and Privacy

2.2 Organisations that develop, make available or use AI systems that surveil human behavior using sensitive personal data (such as data collected in non-public spaces such as the home), facial-recognition data or biometric data shall apply the Transparency and Privacy principles with particular rigour, including as regards the reasonable purpose, limited collection, limited use, limited disclosure and limited retention principles, as well as by providing full transparency as to whether and when a device's voice, movement or image surveillance features have been activated. Sensitive personal data such as biometric data and genetic data collected locally by IoT devices (such as fitness monitors and smart phones) and natural language, movement and image data collected by "always

on" IoT devices (such as personal assistants and smart home devices) shall, to the greatest extent possible, securely store such data, in encrypted format, only locally on the device in a manner that allows for the maximal level of autonomy and control over the data by the individual(s) to whom it relates.

2.3 Organisations that develop, make available or use AI systems that predict and influence human behavior shall put in place appropriate safeguards to promote informed human agency and autonomy and to avoid destructive psychological and behavioural manipulation, addiction, dependency and attention deficit. Such safeguards shall include conducting a responsible AI ethical risk assessment of the technology as part of an accountable governance process prior to deployment of the AI System and ensuring that any such deployment is consistent with respect for other principles of the Policy Framework for Responsible AI such as Transparency and Explainability, Fairness and Non-Discrimination, and Privacy.

3 Work and automation

3.1 Organisations that implement AI systems in the workplace should provide opportunities for affected employees to participate in the decision-making process related to such implementation.

3.2 Consideration should be given as to whether it is achievable from a technological perspective to ensure that all possible occurrences should be pre-decided within an AI system to ensure consistent behaviour. If this is not practicable, organisations developing, deploying or using AI systems should consider at the very least the extent to which they are able to confine the decision outcomes of an AI system to a reasonable, non-aberrant range of responses, taking into account the wider context, the impact of the decision and the moral appropriateness of "weighing the unweighable" such as life vs. life.

3.3 Organisations that develop, make available or use AI systems that have an impact on employment should conduct a Responsible AI Impact Assessment to determine the net effects of such implementation.

3.4 Organisations that develop, make available or use AI systems that surveil or influence employee behavior in the workplace shall put in place appropriate safeguards to promote the informed human agency, autonomy and dignity of employees and to avoid inappropriate or destructive impacts on the emotional or psychological health of employees (monotony of tasks, excessive surveillance, gaming of behavior, continuous exposure to horrific content). Such safeguards shall include conducting a responsible AI ethical risk assessment of the technology as part of an accountable governance process prior to deployment of the AI System and ensuring that any such deployment is consistent with respect for other principles of the Policy Framework for Responsible AI such as Transparency and Explainability, Fairness and Non-Discrimination, and Privacy.

3.5 Governments should closely monitor the progress of AI-driven automation in order to identify the sectors of their economy where human workers are the most affected. Governments should actively solicit and monitor industry, employee and other stakeholder data and commentary regarding the impact of AI systems on the workplace and should develop an open forum for sharing experience and best practices.

3.6 Governments should promote educational policies that equip all children with the skills, knowledge and qualities required by the new economy and that promote life-long learning.

3.7 Governments should encourage the creation of opportunities for adults to learn new useful skills, especially for those displaced by automation.

3.8 Governments should study the viability and advisability of new social welfare and benefit systems to help reduce, where warranted, socio-economic inequality caused by the introduction of AI systems and robotic automation.

4 Environmental impact

4.1 Organisations that develop, make available or use AI systems should assess the overall environmental impact of such AI systems, throughout their implementation, including consumption of resources, energy costs of data storage and processing and the net energy efficiencies or environmental benefits that they may produce. Organisations should seek to promote and implement uses of AI systems with a view to achieving overall carbon neutrality or carbon reduction.

4.2 Governments are encouraged to adjust regulatory regimes and/or promote industry self-regulatory regimes concerning market-entry and/or adoption of AI systems in a way that the possible exposure (in terms of 'opportunities vs. risks') that may result from the public operation of such AI systems is reasonably reflected. Special regimes for intermediary and limited admissions to enable testing and refining of the operation of the AI system can help to expedite the completion of the AI system and improve its safety and reliability.

4.3 In order to ensure and maintain public trust in final human control, governments should consider implementing rules that ensure comprehensive and transparent investigation of such adverse and unanticipated outcomes of AI systems that have occurred through their usage, in particular if these outcomes have lethal or injurious consequences for the humans using such systems. Such investigations should be used for considering adjusting the regulatory framework for AI systems, in particular to develop, where practicable and achievable, a more rounded understanding of how and when such systems should gracefully handover to their human operators in a failure scenario.

4.4 AI has a particular potential to reduce environmentally harmful resource waste and inefficiencies. AI research regarding these objectives should be encouraged. In order to do so, policies must be put in place to ensure the relevant data is accessible and usable in a manner consistent with respect for other principles of the Policy Framework for Responsible AI such as Fairness and Non-Discrimination, Open Data and Fair Competition and Privacy.

5 Weaponised AI

5.1 The use of lethal autonomous weapons systems (LAWS) should respect the principles and standards of and be consistent with international humanitarian law on the use of weapons and wider international human rights law.

5.2 Governments should implement multilateral mechanisms to define, implement and monitor compliance with international agreements regarding the ethical development, use and commerce of LAWS.

5.3 Governments and organisations should refrain from developing, selling or using lethal autonomous weapon systems (LAWS) able to select and engage targets without human control and oversight in all contexts.

5.4 Organisations that develop, make available or use AI systems should inform their employees when they are assigned to projects relating to LAWS.

6 The weaponisation of false or misleading information

6.1 Organisations that develop, make available or use AI systems to filter or promote informational content on internet platforms that is shared or seen by their users should take reasonable

measures, consistent with applicable law, to minimise the spread of false or misleading information where there is a material risk that such false or misleading information might lead to significant harm to individuals, groups or democratic institutions.

6.2 AI has the potential to assist in efficiently and pro-actively identifying (and, where appropriate, suppressing) unlawful content such as hate speech or weaponised false or misleading information. AI research into means of accomplishing these objectives in a manner consistent with freedom of expression should be encouraged.

6.3 Organisations that develop, make available or use AI systems on platforms to filter or promote informational content that is shared or seen by their users should provide a mechanism by which users can flag potentially harmful content in a timely manner.

6.4 Organisations that develop, make available or use AI systems on platforms to filter or promote informational content that is shared or seen by their users should provide a mechanism by which content providers can challenge the removal of such content by such organisations from their network or platform in a timely manner.

6.5 Governments should provide clear guidelines to help organisations that develop, make available or use AI systems on platforms identify prohibited content that respect both the rights to dignity and equality and the right to freedom of expression.

6.6 Courts should remain the ultimate arbiters of lawful content.

Endnotes

1 For another example of an AI governance framework that references the importance of ensuring the protection of human autonomy and human agency, see Singapore's Model Artificial Intelligence Governance Framework, Second Edition: https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf.

2 The Right to Privacy, Samuel D. Warren and Louis D. Brandeis, (1890) 4:5 Harv. L.R. 194.

3 *Ibid.*

4 Privacy tort in general: Th Catanzariti, n. 2 above, 138; W Prosser, "Privacy," (1960) 48 California Law Review 382; H Kalven, "Privacy and tort law—were Warren and Brandeis wrong?," (1966) 31 Law and Contemporary Problems 326; A L Goodhart, "Privacy," (1931) The Law Quaterly Review 23 et seq.

5 US Restatement of Torts (2nd), 1965, § 652A-I.

6 L. Dormehl, "Surveillance on steroids: How A.I. is making Big Brother bigger and brainier," https://www.digitaltrends.com/cool-tech/ai-taking-facial-recognition-next-level/.

7 Y. N. Harari, "The world after coronavirus," https://www.ft.com/content/19d90308-6858-11ea-a3c9-1fe6fedcca75: "Hitherto, when your finger touched the screen of your smartphone and clicked on a link, the government wanted to know what exactly your finger was clicking on. But with coronavirus, the focus of interest shifts. Now the government wants to know the temperature of your finger and the blood-pressure under its skin."

8 B. Marr, "The 10 Best Examples of How AI Is Already Used in Our Everyday Life," https://www.forbes.com/sites/bernardmarr/2019/12/16/the-10-best-examples-of-how-ai-is-already-used-in-our-everyday-life/#1d985c621171.

9 Human Technology Foundation Report: "Technology Governance in Times of Crisis: COVID-19 Related Decision Support," http://optictechnology.org/index.php/en/research, p. 21.

10 UK's House of Lords Library Briefing Note on Predictive and Decision-Making Algorithms in Public Policy, February 2020.

11 N. Powling, "AI: The smart side of surveillance," https://www.computerweekly.com/microscope/opinion/AI-The-smart-side-of-surveillance.

12 S. Feldstein, "The Global Expansion of AI surveillance," https://carnegieendowment.org/2019/09/17/global-expansion-of-ai-surveillance-pub-79847.

13 World Economic Forum, https://www.weforum.org/agenda/2020/04/governments-must-build-trust-in-ai-to-fight-covid-19-here-s-how-they-can-do-it/.

14 https://www.weforum.org/agenda/2020/04/governments-must-build-trust-in-ai-to-fight-covid-19-here-s-how-they-can-do-it/. For a discussion of an approach to the ethical governance of Covid-19 response technologies, see the Human Technology Foundation Report: "Technology Governance in Times of Crisis: COVID-19 Related Decision Support," http://optictechnology.org/index.php/en/research

15 C. Baynes, "Chinese police to use facial recognition software to send jaywalkers instant fines by test," https://www.independent.co.uk/news/world/asia/china-police-facial-recognition-technology-ai-jaywalkers-fines-text-wechat-weibo-cctv-a8279531.html.

16 P. Mozur, "One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority," https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html.

17 "Boston police support the effort to ban facial recognition technology—for now," https://www.boston.com/news/local-news/2020/06/10/boston-facial-recognition-technology-police.

18 "Ethics of AI for Video Surveillance," https://oddity.ai/posts/ethics-of-ai-video-surveillance/.

19 Details of Sapience and Barclays' trial can be seen at https://www.bbc.co.uk/news/explainers-51571684.

20 An example of this are the 2019 revelations that Amazon employees could hear private information recorded by Amazon's Alexa system, https://www.theguardian.com/technology/2019/apr/11/amazon-staff-listen-to-customers-alexa-recordings-report-says.

21 Z. Villnes, "Watch Out: The Psychological Effects of Mass Surveillance," https://www.goodtherapy.org/blog/watch-out-psychological-effects-of-mass-surveillance-0910137.

22 C. Chambers, "NSA and GHCQ: the flawed psychology of government mass surveillance," https://www.theguardian.com/science/head-quarters/2013/aug/26/nsa-gchq-psychology-government-mass-surveillance.

23 J. Penney, "Chilling Effects: Online Surveillance and Wikipedia Use," Berkeley Technology Law Journal, Vol. 31, No. 1, p. 117, 2016.

24 D. Lyon, *The Culture of Surveillance: Watching as a Way of Life*, Cambridge, Polity Press, 2018.

25 Z. Villnes, "Watch Out: The Psychological Effects of Mass Surveillance," https://www.goodtherapy.org/blog/watch-out-psychological-effects-of-mass-surveillance-0910137.

26 High-Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy AI," 2019, 12.

27 For an in-depth discussion of COVI app, see Human Technology Foundation Report: "Technology Governance in Times of Crisis: COVID-19 Related Decision Support," http://opticteчnology.org/index.php/en/research, pp. 108-113. See also, Alsdurf, Belliveau, Bengio et al, "COVI White Paper—Version 1.1," July 27, 2020, arxiv, https://arxiv.org/abs/2005.08502.

28 Zuboff, Shoshana, Age of Surveillance Capitalism: The fight for a human future at the new frontier of power, Published 2019. Rather than read the tome of 664 pages, here is a summary: https://www.theguardian.com/books/2019/feb/02/age-of-surveillance-capitalism-shoshana-zuboff-review.

29 https://privacyinternational.org/learning-topics/data-and-elections.

30 Age of Surveillance Capitalism: The fight for a human future at the new frontier of power, Shoshana Zuboff, published 2019.

31 Arguably Big Data analytics is not an activity regarding just one individual's data sets, but that of a group of individuals. Current data protection laws tend to follow an individual-oriented model which is less able to fully acknowledge the novelty and complexity of data formed from a group. Greater regard should therefore be had to the rights of the group, see: Group Privacy: New Challenges of Data Technologies Linnet Taylor, Luciano Floridi, and Bart van der Sloot, 2017.

32 Snowden, Edward, Permanent Record (September 2019).

33 Sunstein, Cass R., Misconceptions About Nudges (September 6, 2017): https://ssrn.com/abstract=3033101.

34 Snowden, Edward, Permanent Record (September 2019) p. 172.

35 Graystone, Andrew, Too Much Information? (2019).

36 https://www.wired.com/story/tristan-harris-tech-is-downgrading-humans-time-to-fight-back/.

37 Wu, Tim, Blind Spot: The Attention Economy and the Law (March 2017). Antitrust Law Journal, https://ssrn.com/abstract=2941094.

38 Wu, Tim, *The Attention Merchants: The Epic Scramble to Get Inside Our Heads*.

39 Tristan Harris quoted as saying "Inadvertently, whether they want to or not, they are shaping the thoughts and feelings and actions of people. They are programming people." https://www.cbsnews.com/news/brain-hacking-tech-insiders-60-minutes/.

40 *Supra.*

41 "Brain hacking," a concept so named by Tristan Harris (ex Google) can result in attention deficit, addiction and coerced behaviour. https://www.cbsnews.com/news/brain-hacking-tech-insiders-60-minutes/.

42 Religious rights and choice under the European Convention on Human Rights, Peter W. Edge, 2000, 3 Web Journal of Current Legal Issues with reference to the case of: *Kokkinakis v Greece* (1994) 17 EHRR 397.

43 Weinmann, Markus and Schneider, Christoph and vom Brocke, Jan, Digital Nudging (2015). Weinmann, M., Schneider, C. & vom Brocke, J. (2016). Digital Nudging. Business & Information Systems Engineering, 58(6): 433-436. https://ssrn.com/abstract=2708250.

44 Balkin, Jack M., Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation (September 9, 2017). UC Davis Law Review, (2018); Yale Law School, Public Law Research Paper No. 615: https://ssrn.com/abstract=3038939.

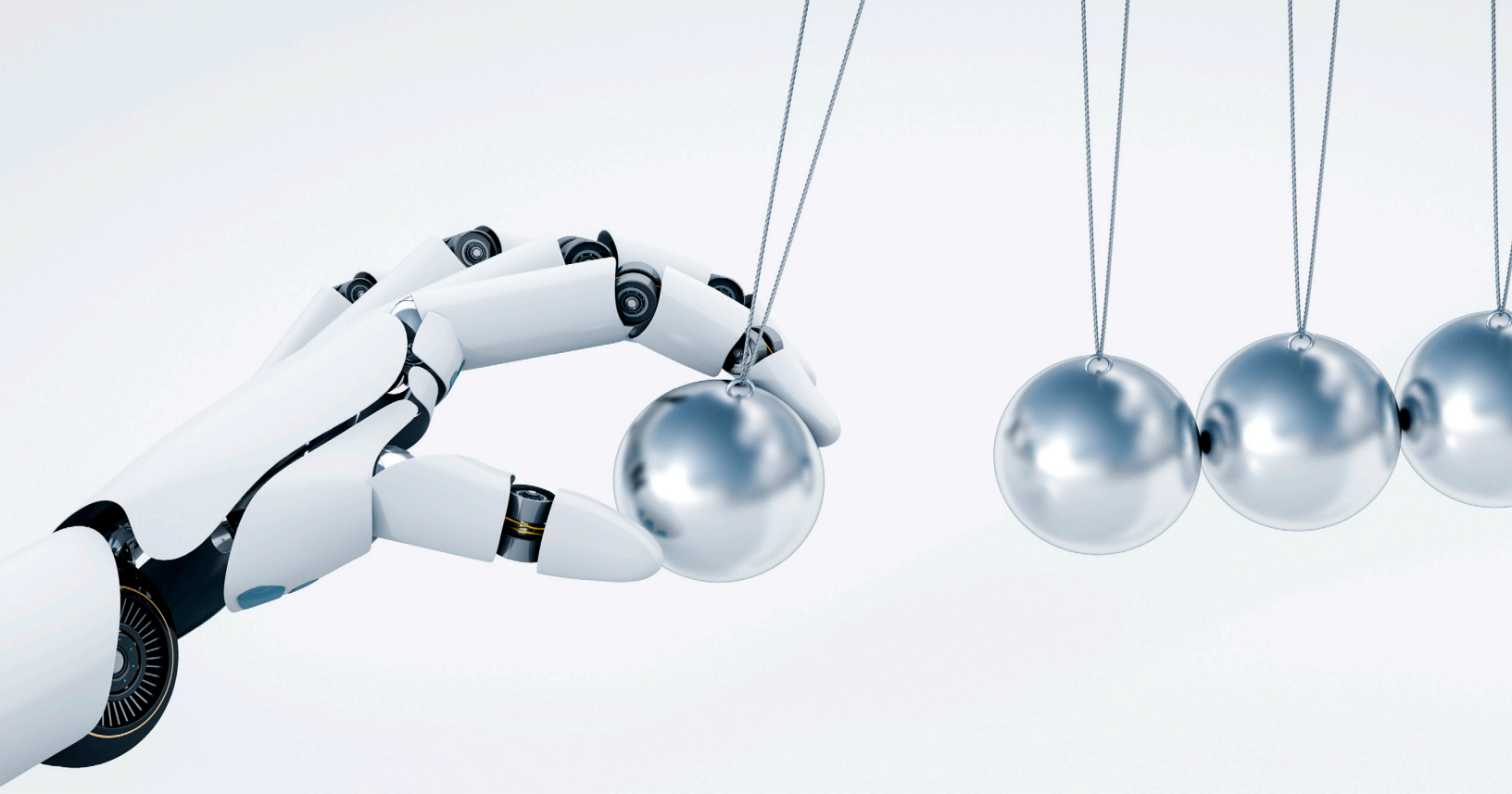
45 Steven, Matthew, The Inviolate Person, January 2017, p. 23.

46 http://alanwinfield.blogspot.com/2019/06/energy-and-exploitation-ais-dirty.html.

47 https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.

48 https://www.theguardian.com/technology/2019/sep/17/revealed-catastrophic-effects-working-facebook-moderator.

49 https://www.washingtonpost.com/technology/2019/07/25/social-media-companies-are-outsourcing-their-dirty-work-philippines-generation-workers-is-paying-price/.



Principle 2

ACCOUNTABILITY

Responsible AI

A GLOBAL POLICY FRAMEWORK

Principle 2 Commentary

ACCOUNTABILITY

Nikhil Narendran | Partner, Trilegal

Patricia Shaw | CEO and Founder, Beyond Reach Consulting Limited

Introduction

Since the publication of the first edition of *Responsible AI,* the use of AI in systems in sectors of critical importance such as health and finance sectors has increased significantly, not least (in the former case) due to the impact of the COVID-19 pandemic. Notwithstanding the foregoing, with the growth of the use of AI-based systems on a large scale both by corporations and governments, there is a corresponding need that arises to ensure that users and other individuals impacted by such systems are accorded adequate protections in the design and operation of the AI.

This almost certainly will have a notable impact on individuals in ways affecting their quality of life and livelihoods. We have therefore taken this as a sign that our AI principle of accountability must be reviewed to provide additional safeguards in consideration of the individual users or persons impacted by the AI's decisions and actions. Any failure to do so may greatly reduce the ability of such affected persons to take steps to protect their rights, freedoms and status. It is also acknowledged however that standards of accountability cannot be applied in an equal manner across all types of AI systems, but must take note of the extent of the impact which may be caused by its use which in turn will determine the appropriate accountability actions to be taken.

Principle 2 has consequently been revised to highlight the requirement to conduct a Responsible AI Impact Assessment and identify an accountable person, particularly in cases where the degree of autonomy and criticality is high.

It is however advisable that a Responsible AI Impact Assessment also be conducted in respect of all AI projects and at various intersections of the AI lifecycle in order to determine levels of risk and capture any changes to the risk that may transpire during the course of the AI lifecycle.

it is recognised that unless a cohesive move from international legislators in respect of both law, ethical principles and standards of governance to hold AI to account, both governments and private organisations alike will continue to apply globally inconsistent regulatory approaches. Discussed in further detail below, this has led us to update principle 2 recommending governments to work together in a more collaborated and coordinated manner, and to seek to identify and address accountability gaps in existing legal and regulatory frameworks applicable to AI systems.

Accountability in general—legal background

The European Commission recently observed that, *"The self-learning feature of the AI products and systems may enable the machine to take decisions that deviate from what was initially intended by the producers and consequently what is expected by the users. This raises questions about human control, so that humans could choose how and whether to delegate decisions to AI products and systems, to accomplish human-chosen objectives."*[1]

Human oversight is a critical, and we would argue an, essential, feature of any functionally effective accountability mechanism.

The Ethics Guidelines for Trustworthy AI prepared by the High-Level Expert Group on Artificial intelligence set up by the European Commission and made public on 9 April 2019 (**HLEG Guidelines**) as well as the Model Artificial Intelligence Governance Framework: Second Edition released by Singapore in 2020 (**Singapore Framework**) have identified three mechanisms of human oversight involving differing levels of human control over the AI systems, albeit with different nomenclature.

The Singapore Framework highlights that *"For safety-critical systems, it would be prudent for organisations to ensure that a person be allowed to assume control, with the AI system providing sufficient information for that person to make meaningful decisions or to safely shut down the system where human control is not possible."*[2] Internal transparency and explainability across organisational functions being a fundamental pre-requisite.

According to the HLEG Guidelines, *"human oversight helps ensure that an AI system does not undermine human autonomy or causes other adverse effects. Oversight may be achieved through governance mechanisms such as a human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approach."*[3] The three oversight mechanisms discussed are as follows:

- Human-in-the-loop, where humans retain full control to intervene in every decision-making cycle. This concept is the same in the Singapore Framework.
- Human-on-the-loop, where humans can intervene during the design cycle of the system and may carry out monitoring. This is similar to the human out of the loop concept in the Singapore Framework, where there is no human override option or oversight over the execution of decisions.
- Human-in-command, where humans can oversee the overall activity of the AI system and decide the situations and manner in which it may be used. This concept is similar to the human over the loop model envisioned by the Singapore Framework.

While implementing the above models, it is critical to identify the 'human behind the machine,' particularly in those cases where the AI system carries out a greater portion of decision-making and has greater autonomy. Due to the limited intervention possible in some of these models, there is a risk that the AI system will function without taking into account the harms caused to affected individuals from the decision-making, notably in cases where the criticality of the system is higher.

To provide for a recourse and a clearer attribution of liability, it is important to have transparent systems allowing independent third parties (whether they are regulators, those auditing the AI system, those investigating the outcomes of an AI system, or the end users themselves) to inquire of the organisation, and for the organisation to be able to identify the named individual(s) or organisational role(s) behind the systems holding internal organisational and/or external public facing responsibility for (a) the good and proper functioning of the AI system in line with the organisation's intended outcomes and (b) providing the organisation with assurance that the AI system is legally (and potentially also ethically) compliant and has appropriate risk management and/or mitigation measures in place to abate poor or unlawful outcomes, inappropriate unintended consequences, and harm to individuals or people groups. This has been introduced as an additional concept in principle 1.4. However, we have also made changes to principle 1 to reflect the core principle of keeping humans behind the machines, maintaining human centred AI with the machine-in-the-loop.[4]

Given that some of the additional accountability measures may be specific to industries or sectors, it was felt that appropriate weight be given to sector specific accountability measures. For instance, a decision in the healthcare sector has an impact on life and therefore, the sector may consider additional safeguards (compared to another industry such as lifestyle sector), for example in case of autonomous healthcare applications which has an interplay with social welfare insurance and healthcare systems. This has also been reflected in principle 1. Throughout the discussion in *Responsible AI: A Global Policy Framework* we have relied on the foundation that the accountability principle is interconnected with our other defined principles. For example, principle 1 challenges organisations to be accountable for risks beyond that merely of legal compliance considering the ethical and societal outcomes of the AI systems those organisations produce; in order for principle 2 to properly hold organisations to account requires both internal and external transparency and explainability as outlined in principle 3; to attribute responsibility and to hold an organisation (and the individuals behind them) to account for their actions/inactions resulting in poor outcomes and harms produced by the AI system requires an organisation to have a clear understanding of: (i) what AI systems outcomes are fair, unfair, biased and/or unlawful and discriminatory in the particular context of that organisation's AI system in accordance with principle 4, (ii) what is safe and reliable in accordance with principle 5, (iii) how the use of open data and competitive practices impacts an organisation's AI system in accordance with principle 6, (iv) how AI systems impact on individual and group privacy and can either enhance it or be encroach upon it in accordance with principle 7, and (v) how ownership of AI helps identify which organisation or individual(s) should be held accountable but also apportionment of that responsibility in accordance with principle 8. This was not adequately reflected in our prior draft of principle 1 and therefore we have incorporated it accordingly.

In cases where the extent of autonomy of the AI system and its criticality are both high, it becomes even more critical to identify the person involved and accountable for its function. Fundamentally there must always be a legal person who is held to account, to help lend legitimacy to the operation, provide a clear source of authority over its functioning, provide an element of justification for the internal decision, and provide for a point of contact if and when users seek recourse for any grievances or harms to them arising from the AI system. However, identifying the accountable person(s) may be practically problematic. In many situations, it may become difficult to pin-point a distinct responsible organisation or role with which responsibility should reside for a given element of the AI system (although accountability and therefore liability will likely be apportioned contractually). According to the European Commission's "Report

on the safety and liability implications of Artificial Intelligence, the Internet of Things and Robotics" dated 19 February 2020, liability should reside with the Operator, and the prevailing direction within the EU is likely to be consistent with EU product liability legislation, namely on a strict liability basis.

Such organisations will therefore need to have clearly defined roles and responsibilities set prior to rolling out AI projects and for the ongoing management and monitoring of AI post deployment up to point of sunsetting to ensure accountability is clear and apportioned fairly in accordance with the organisation best positioned to do it.

Responsible AI Impact Assessments

As we mention above, we strongly recommend undertaking a Responsible AI Impact Assessment where critical or high risk AI is implemented. Determining what amounts to high risk AI is difficult. For guidance on what is deemed "high risk," the European Commission in its White Paper on AI is of the view that an AI application may be considered high risk where is satisfies two criteria:

> *"First, the AI application is employed in a sector where, given the characteristics of the activities typically undertaken, significant risks can be expected to occur. This first criterion ensures that the regulatory intervention is targeted on the areas where, generally speaking, risks are deemed most likely to occur. The sectors covered should be specifically and exhaustively listed in the new regulatory framework. For instance, healthcare; transport; energy and parts of the public sector. The list should be periodically reviewed and amended where necessary in function of relevant developments in practice;*
>
> *Second, the AI application in the sector in question is, in addition, used in such a manner that significant risks are likely to arise. This second criterion reflects the acknowledgment that not every use of AI in the selected sectors necessarily involves significant risks. For example, whilst healthcare generally may well be a relevant sector, a flaw in the appointment scheduling system in a hospital will normally not pose risks of such significance as to justify legislative intervention. The assessment of the level of risk of a given use could be based on the impact on the affected parties. For instance, uses of AI applications that produce legal or similarly significant effects for the rights of an individual or a company; that pose risk of injury, death or significant material or immaterial damage; that produce effects that cannot reasonably be avoided by individuals or legal entities."*[5]

Fundamentally, for accountability to work in any organisation there needs to be robust governance and oversight, and in this regard we would strongly advocate the need for "humans behind the machines," as discussed above.

Since the publication of the first edition of *Responsible AI: A Global Policy Framework*, a great deal of discussion (both academic and practitioner led) has been had about a global coordinated and collaborative approach to AI regulation and AI Governance.[6]

To regulate requires both legislators and regulators to have *"developed sufficiently comprehensive expertise to formulate standards that reflect not only the technological and engineering perspectives but also*

legal and ethical considerations"[7] and that any such laws would have to be to some degree high-level and technology neutral, perhaps referencing more dynamic international standards as a way of keeping technologically and culturally relevant and futureproof.

It is recognised that without a cohesive move from international legislators in respect of both law, ethical principles and standards of governance to hold AI to account, both governments and private organisations alike will continue to apply globally inconsistent regulatory approaches.

To that end, Principle 2 has been updated to recommend that Governments *should seek to work collaboratively and in a coordinated manner across the international landscape to apply the principles of this Policy Framework for Responsible AI or other analogous internationally recognised principles to ensure consistency of approach and application when holding AI systems to account.*

Furthermore, given that each jurisdiction has law applicable to AI and its impacts, and that these may address accountability and provide for enforcement and redress in and through different means, Principle 2 has been further updated to recommend that accountability gaps should be identified and addressed. It is recognised that this will not be an easy feat, given that many governments and extra-governmental organisations are seeking to promote innovation and obtain the many economic benefits of AI, whilst not stifling that innovation and mitigating against the individual and societal harms both in the short and longer term.

Finally, there is now an opportunity for better law-making in the realm of AI by intertwining it with globally recognised standards including the requirement to ex-ante and ex-post risk assess[8] (with the potential for inclusion of certification or licensing models of regulatory operations).

Whether it is hard law or soft law that ultimately achieves it, it is the pressure of legal or non-legal sanctions that act as an enabler of consistent accountability of AI. As seen from the overarching principles laid out in this *Responsible AI: A Global Policy Framework* or other analogous internationally recognised principles, principles alone cannot achieve consistent accountability without the role of law. Law can empower and incentivise accountability through enforcement, penalising inappropriate unethical behaviours, and provide reasonable and proportionate redress (and potentially compensation) for those who have been harmed by AI.

Principle 2
Accountability

Organisations that develop, make available or use AI systems ought to be accountable for the consequences of their actions and shall designate an individual or individuals who are accountable for the organisation's compliance with the principles of the Policy Framework for Responsible AI or other adopted principles (including analogous principles that may be developed for a specific industry) with the objective of keeping humans behind the machines and AI Human centric.

1 Accountability

1.1. The identity of the individual(s) designated by the organisation to oversee the organisation's compliance with the principles shall be made known upon request.

1.2. Organisations that develop, make available deploy or use AI systems shall use human oversight to carry out determination of the situations in which to carry out delegation to AI decision-making, while ensuring that such use is to accomplish human-chosen objectives. Human oversight can be achieved through three mechanisms, i.e. human-in-the-loop (where humans retain full control to intervene in every decision-making cycle), human-on-the-loop (where humans can intervene during the design cycle of the system and may carry out monitoring) and human-in-command (where humans can oversee the overall activity of the AI system and decide the situations and manner in which it may be used).

1.3. Organisations that develop, make available deploy or use AI systems shall implement policies and practices to give effect to the principles of the Policy Framework for Responsible AI or other adopted principles (including analogous principles that may be developed for a specific industry), including:

i. establishing processes to determine whether, when and how to implement a "Responsible AI Impact Assessment" process;

ii. establishing and implementing "Responsible AI by Design" principles;

iii. establishing procedures to receive and respond to complaints and inquiries;

iv. training staff and communicating to staff information about the organisation's principles, policies and practices; and

v. developing information to explain the organisation's principles, policies and procedures.

2 Government

2.1. Governments should seek to work collaboratively and in a coordinated manner across the international landscape to apply the principles of this Policy Framework for Responsible AI or other analogous internationally recognised principles to ensure consistency of approach and application when holding AI systems to account.

2.2. Governments that assess the potential for "accountability gaps" in existing legal and regulatory frameworks applicable to AI systems

should adopt a balanced approach that encourages innovation while mitigating against the risk of significant individual or societal harm.

2.3. Any such legal and regulatory frameworks should promote the eight principles of the Policy Framework for Responsible AI or encompass similar considerations and consider appropriate legal and regulatory enforcement and redress mechanisms.

2.4. Governments should not grant distinct legal personality to AI systems, as doing so would undermine the fundamental principle that humans should ultimately remain accountable for the acts and omissions of AI systems.

2.5. Governments should be transparent and put appropriate human oversight mechanisms in place when utilising AI systems for products or services which are in the public interest, and ensure that the objective and outcomes of such AI Systems are understood by its subjects or citizens.

3 Contextual approach

3.1. The intensity of the accountability obligation will vary according to the degree of autonomy and criticality of the AI system and its potential to cause individual or societal harm. The greater the level of autonomy of the AI system and the greater the criticality of the outcomes that it may produce, the higher the degree of accountability that will apply to the organisation that develops, deploys or uses the AI system ("High Risk AI").

3.2. Where an AI system is deemed to be High Risk AI, a Responsible AI Impact Assessment ("RAIIA") should be conducted and clearly identify the accountable person(s).

Endnotes

1 "Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and Robotics" dated 19 February 2020, p. 7.

2 Singapore Framework, p. 31.

3 HLEG Guidelines, p. 16.

4 Human Centred machine learning: machine in the loop approach—https://medium.com/@ChenhaoTan/human-centered-machine-learning-a-machine-in-the-loop-approach-ed024db34fe7.

5 White Paper on Artificial Intelligence—A European Approach to Excellence and Trust, dated 19 February 2020, issued by the European Commission, p. 17.

6 See Orbit Conference Paper 100+ Women in AI Ethics 2019, https://khcdn8dab2d6280.b-cdn.net/wp-content/uploads/2020/03/100-Brilliant-Women-2019-conference-report.pdf.

7 Council of Europe Study, Algorithms_and_Human_Rights, 2017.

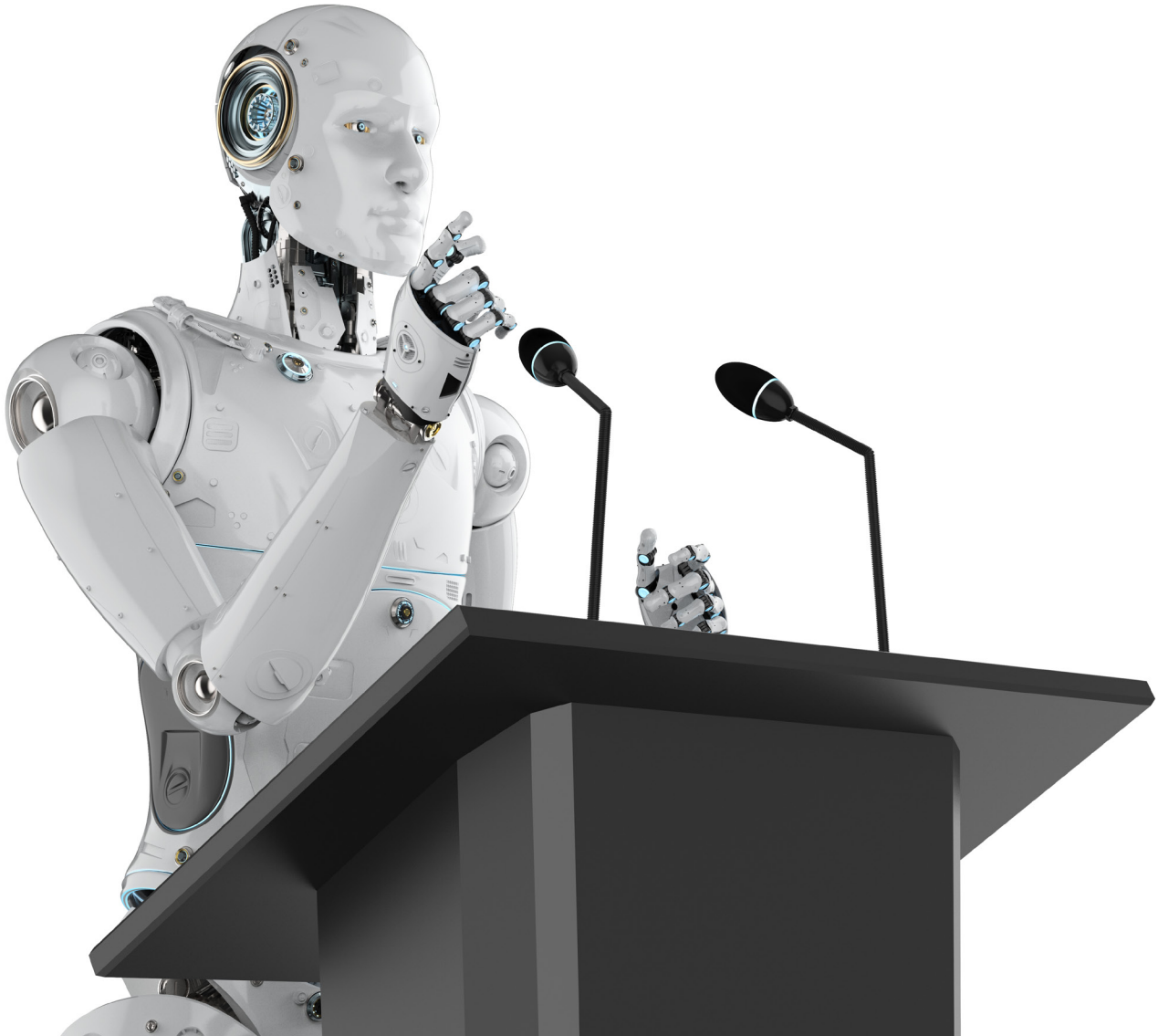
8 Fosch Villaronga, Eduard, and Golia, Angelo Jr., Robots, Standards and the Law: Rivalries between private standards and public policymaking for robot governance, Computer Law & Security Review, January 2019, DOI: 10.1016/j.clsr.2018.12.009.

Responsible AI

A GLOBAL POLICY FRAMEWORK

Principle 3

TRANSPARENCY AND EXPLAINABILITY



TRANSPARENCY AND EXPLAINABILITY

CHAPTER LEAD
Kevin Fumai | Oracle America, Inc., United States

Richard Austin | Deeth Williams Wall LLP, Canada

Carmen De la Cruz Böhringer | de la cruz beranek Rechtsanwälte AG, Switzerland

Charles Morgan | McCarthy Tétrault, Canada

Ursula Widmer | Dr. Widmer & Partners, Switzerland

Anthony Wong | AGW Consulting Pty Ltd, Australia

In the first edition of *Responsible AI: A Global Policy Framework*, we argued that promoting transparency and explainability was a critical next step in AI's evolution. We remain convinced of that necessity with the passage of time, drawing strength from a chorus of voices from public and private communities advocating for these fundamental values. However, we have observed that few recommendations have been made to translate these high-level principles into practice. We have therefore focused this update on how transparency and explainability can become more integrated within the AI governance model.

Refocusing the target

Our first update turns on the threshold issue of to whom the principle applies. In the original publication, we suggested binding all organisations that "develop, deploy or use" AI systems, deeming the phrase sufficiently broad to capture the key actors in the AI ecosystem. We now feel a more nuanced approach is appropriate based on the emerging realities of the AI marketplace.

In a February 2020 white paper, the European Commission recognised that traditional safety laws were designed to apply to the organisation that brings a product to market, but found such an approach can ignore the unique ways AI can be bundled into new products and services by organisations other than the original AI developer.[1] The Commission indicated that these laws should be better tailored to the disparate actors involved in the AI lifecycle, placing obligations on those best positioned to address potential risks. For example, while an AI developer may be able to manage risks arising from the initial development phase, its ability to control risks during a downstream use of the system by an independent organisation may be more limited. In that case, the organisation using the AI system should bear more responsibility since it designs the parameters of the use case and data strategy, and incorporates the output from the AI system into its own business offerings.[2]

We agree with this analysis and consider it consistent with the basic accountability framework we expressed in the original publication. However, to better capture the groups of actors that are involved in

the AI lifecycle, throughout the updated version of the *Responsible AI* framework, we have replaced the phrase "develop, deploy or use" with "develop, make available or use."

The rationale for this enhancement is twofold. First, there is an overlap between "deploy" and "use" from an etymological perspective, and that overlap ignores the scenario mentioned above when AI systems are utilized by downstream organisations for their own purposes. Second, the phrase "make available" is a well-known and understood legal concept—particularly in the context of copyright law—so adopting it will provide better clarity and predictability.

Accordingly, this principle will now apply to the following categories of AI participants, while recognizing the possible distinctions among them: (a) those that develop an AI system; (b) those that make an AI system available to third parties, including through a cloud computing platform; and (c) those that use an AI system.

Recognising industry standards

Our second update addresses the other focal point of the Transparency and Explainability Principle: national laws that govern the use of AI systems.

There has always been a delicate balance between innovation and its potential for societal advancement on one hand, and regulation and its need to protect those who may be left behind or harmed on the other. After publishing our first edition, we realized that establishing an equilibrium for responsible AI is not two-dimensional. Instead, industry standards must also be taken into consideration because they complement national laws and fill accountability gaps.[3]

The process by which industry standards are created is comprehensive and inclusive, with companies, consumers, academia, and governments as contributing participants. This consensus-driven approach has enabled industry standards to influence the development and use of new technologies even though they do not have the force of law.

The initial efforts to create AI standards offer promise for transparency and explainability. For example, the Institute for Electrical and Electronics Engineers ("IEEE") released a report entitled *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems* that, among other things, calls for transparency both in favour of individuals affected by an AI system and the public at large.[4] We have supported this same broad scope of disclosure in this update, on the basis that transparency is one of the best disinfectants to the potential risks of AI.

Similarly, the United States National Institute of Standards and Technology ("NIST") recently released a draft publication entitled *Four Principles of Explainable Artificial Intelligence* that presents five types of explanations an AI system may offer.[5] NIST also published another report entitled *A Plan for Federal Engagement in Developing Technical Standards and Related Tools* that provides guidance on, and establishes requirements for, explainability, including with a detailed list of benchmarks, metrics, and open source software and data repositories.[6] Captured within that list is another project that deserves special

attention: the Explainable AI Program by the United States Defense Advanced Research Projects Agency ("DARPA"), which has made available a library of tools and techniques that promote explainability during the development lifecycle of every major AI domain.[7]

This sampling only begins to tease the potential of AI industry standards. Indeed, because we recognise their early value and expect it to only increase over time, we have incorporated a reference to industry standards within all eight principles of our Responsible AI framework.

Detailing the obligations

The need for more specific, actionable guidance on transparency and explainability has risen as AI continues to permeate everyday life. This section sets forth the criteria and rationale we propose for each to guide organisations that use or make an AI system available in a decision-making process.

To meet the transparency obligation, we recommend those organisations should disclose "meaningful information" on four main topics, namely: (1) the fact that an AI system is being used in a decision-making process; (2) the intended purpose(s); (3) the types of data sets that are used and generated by the AI system; and (4) whether and to what extent the decision-making process may include human participation.

The first two of these topics combine to set the baseline of whether and for what purpose(s) an AI system may make a decision affecting an individual's rights. It is essential for individuals to know that an AI system is involved and how the system will be used if they are to make an informed decision about whether they wish to proceed, and, if so, for them to be armed with a better understanding of the interaction and how it may affect them. That knowledge, in turn, will help establish the rights and remedies that should protect individuals who choose to interact with the AI system.

The third topic relates to the data used and generated by the AI system. All data have value, yet some data have more value and are more sensitive than others. In particular, the presence of personal information informs the gradual and contextual approach discussed in Section 6 below, and links to the requirements of the Privacy Principle for protecting such information in compliance with applicable laws and regulations.

The final topic concerns whether a human is involved in the AI decision-making process. Despite the continuing advances in AI, the majority of algorithms require human participation in some capacity and, in our opinion, these algorithms will hold the most promise when they are designed with a "human-in-the-loop" capability that enables human judgment to augment, or be augmented by, machine prediction. The disclosure of this information is designed to address that reality, creating a feedback loop to the first transparency topic in the process—i.e. when an individual is advised that an AI system is used with human interaction (or not), they can make a more informed decision on whether to continue.

Collectively, meaningful disclosure on these topics sets a detailed standard for transparency in our Responsible AI framework, ensuring the fundamental human privileges of choice and control are not wrested away by AI systems.

Our proposed disclosure obligations for explainability are also based on four additional topics, namely: (1) the transparency information discussed immediately above; (2) information that offers meaningful interpretability of the algorithmic logic of the AI system; (3) meaningful information to understand the decision/outcome; and (4) information regarding how the individual may contest the decision or outcome.

The first disclosure obligation, while seemingly redundant, borrows its logic from the privacy domain. Specifically, there is a truism in privacy that you cannot have privacy without security, but you can have security without privacy. The form of that truism applies readily here—i.e. you cannot have explainability without transparency, but you can have transparency without explainability. It is for this reason that our discussion on explainability begins, but does not end, with the transparency requirements described above.

The next two topics establish the core of the explainability obligation, aiming to help individuals understand what algorithmic logic and factors were taken into consideration by an AI system and were material to an outcome. This information can be presented in different ways, including, for example, through visual interfaces or interactive tools that show how a change in one piece of information would have led to a different decision (a "counterfactual"),[8] and scoring or saliency models that identify and weight factors relevant to a decision. Regardless of the presentation style, this information is critical to engender trust and provide a basis to challenge a decision—and is often legally required, as in the case, for example, when a lender declines a loan application in the United States.

These two topics also tease one of the main perceived deterrents to explainability—i.e. that more understanding of an AI system will result in less accuracy or decreased performance. We agree with the emerging consensus (e.g. from DARPA) that this is a technological red herring which ignores the advances in modeling and algorithmic design that can establish a more appropriate balance, even in the most complex AI domains like deep learning. However, we recognize that disclosing too much information creates a different, truly legitimate risk: it can make an AI system more susceptible to manipulation or attack from malicious actors. This is another reason for the gradual, contextual approach outlined in Section 6 below.

Our final topic relates to how an individual may contest the outcome from an AI system. Although we expect AI to continue improving over time, like human decision-making, it will remain imperfect. Organisations will more easily build trust—and make greater use of AI for societal benefit—if they establish and maintain mechanisms to continuously evaluate AI systems and decisions. And, of course, an important part of any such mechanism would be to inform individuals that they have rights to withdraw or dispute a decision.

Only with all these categories of information that promote transparency and explainability will our human-centric values shine through the opacity of AI systems.

Adopting an ex ante, post facto approach

Armed with these new obligations, we turn to the question of when the information must be shared by organisations that use or make an AI system available in a decision-making process. The simple answer that springs to mind—before the AI system is used—ignores the nuanced analysis that must accompany this new level of detail.

Again borrowing from the privacy realm, notice is only effective if given prior to an event happening. In that case, an informed decision on whether personal information may be collected and used can only be made prior to such collection or use. With AI, the transparency obligations defined above will only serve their intended purpose if the information is made available to an affected individual *before* an AI system is used.

Explainability is different, as it requires a more holistic disclosure, both before the AI system is used (to satisfy the transparency requirement) and after. Indeed, because each person is unique and each decision should be as well, an explanation of a decision that has not yet been made is simply not possible.[9] Therefore, we recommend that the remaining information to establish explainability be disclosed "promptly after" an AI decision is made to ensure an affected individual has the ability to asses whether it was justifiable and, if not, the opportunity to appeal.

Removing the "state of the art" qualification

We decided it was also prudent to prune the principle in this update. Specifically, we have removed the caveat that transparency and explainability should depend upon "the state of the art of the technology."

As mentioned in Section 8 below, transparency and explainability form the basis for the entire Responsible AI framework. They allow individuals to know when an AI system is involved, how it is being used, and how it makes a decision, all of which create the foothold to grant fundamental, human-centric rights. Allowing a back door tied to an amorphous "state of the art" catchall could potentially create an escape hatch or slippery slope that undermines the entire framework's utility.

In our view, these obligations should apply regardless of the AI technology being used.[10]

Establishing a gradual, contextual scale

Context always matters, and what may be sufficient in one situation may be deficient in another. We therefore believe that the intensity of the transparency and explainability obligations should depend upon the circumstances. This belief would be hollow, though, without guidance on the factors that should be taken in consideration when evaluating the intensity of the disclosure obligations. We have identified four factors that are particularly relevant when assessing the intensity of these obligations.

First, the scale should consider whether the individual affected by the AI decision is a lay person or expert, as the breadth and depth of information given to the former would be materially different from that given

to the latter. To account for this variance, we recommend adopting a layered disclosure approach or a reasonable person standard.

Second, the scale should contemplate the organization using or making the AI system available, primarily to distinguish between private and public sector organizations. In particular, governmental use of AI may implicate constitutional or other legal rights requiring heightened disclosure.

Third, the disclosure intensity should account for whether sensitive data is used by the AI system or significant legal or human rights are affected. Intuitively, disclosure will usually be higher where sensitive personal data is used and where the outcome of the decision will have a material impact on the affected individual's legal or human rights or similarly significant interests (e.g. when an AI system is used to screen job applicants). Conversely, the disclosure intensity will usually be lower where non-sensitive personal data or de-personalised data is used, or where the impact on the affected individual's legal or human rights or similarly significant interests are relatively inconsequential (e.g. when an AI system suggests music based on previous downloads).

In essence, our general rule is that the higher the level of potential harm, the higher the disclosure obligations. We adopt this logic in Section 4.4 of the principle, requiring that organisations also provide meaningful information in "high intensity" situations on the: (a) traceability and auditability of the algorithmic logic of the AI system; and (b) testing methods used to promote the principles within this policy framework. These additional categories are key inputs necessary to establish accountability, and their inclusion helps anchor the "high intensity" endpoint of the gradual, contextual scale.

Embedding transparency and explainability by design

Design thinking is often used to address complex problems, such as those affecting privacy and data protection.[11] We think a similar approach should be applied to AI, and specifically advocate for transparency and explainability by design in Section 5.1 of the principle.

In particular, we argue that organisations that develop AI systems should ensure the system architecture, algorithmic logic, data sets, testing methods, and all related development and operational policies and procedures default to embed transparency and explainability by design. We recommend that this framework be implemented from the outset and maintained throughout the development and governance lifecycles to promote transparency and explainability that complements the intended purposes of the AI system.

While most think of a "by design" approach as minimising the costs of a problem, we deem explainability by design as also offering its own benefit—i.e. it creates opportunities to improve AI system performance. More specifically, every AI system is trained to learn rules and operate based on its test parameters and data. But no matter the scope of training, there will inevitably be edge cases once an AI system is deployed that fall at or outside the scope of the learned function, and embedding explainability into the governance lifecycle will enable those instances (and future ones) to be better addressed. Essentially, AI developers

can achieve a true win-win outcome when considering explainability as a full-sum complement to AI's predictive capabilities.[12]

Branching transparency and explainability to another principle

As mentioned above, transparency and explainability form the bedrock foundation for trustworthy AI, and our first edition reflected this paradigm through the connective tissue that links this principle to four others—i.e. Accountability, Fairness and Non-Discrimination, Safety and Reliability, and Privacy. We take the occasion of this chapter update to bridge transparency and explainability to the goals of human autonomy and human agency, and, accordingly, have added a link to the Ethical Purpose and Societal Benefit Principle in Section 1.2 of our updated principle.

In the AI context, human autonomy can be characterized as "freedom from subordination to, or coercion by, AI systems."[13] This freedom is only possible through choice, which, in turn, only follows when individuals are provided transparency on whether and to what extent an AI system may be used in a decision-making process. Put simply, without the benefit of AI transparency, individuals may experience a loss of control over their lives.

Human agency ponders the question of what it means to be human. While this may seem too philosophical in the abstract, it is nonetheless critical in today's world with AI systems becoming ubiquitous in our personal and professional lives. Transparency and explainability enable individuals to understand the role of these AI systems and ensure free will and other fundamental rights may be exercised.

■ ■ ■

With the changes outlined above, we believe that the transparency and explainability principle in our Responsible AI framework will begin to translate principle into practice, helping to unlock the capabilities of AI systems to augment human potential. However, we recognize that this journey toward transparency and explainability could be like a ship at sea that will never reach a port unless these principles continue to evolve and are truly and consistently applied in practice on a global scale.

Principle 3
Transparency and Explainability

Organisations that develop, make available or use AI systems, and any national laws or industry standards that govern such use, shall ensure that such use is transparent and that the decision outcomes of the AI system are explainable.

1 Purpose

1.1 The Transparency and Explainability principle aims to promote and maintain public trust in AI systems by requiring organisations that develop, make available and use AI systems to provide sufficient information to demonstrate whether decisions made by the AI systems are fair and impartial, support human agency and human autonomy and establish meaningful responsibility and accountability of an AI system's developers and users.

1.2 The Transparency and Explainability principle supports the Ethical Purpose and Societal Benefit principle, the Accountability principle, the Fairness and Non-Discrimination principle, the Safety and Reliability principle and the Privacy principle.

2 Transparency

2.1 Organisations that make available or use an AI system in decision-making processes which produce legal effects concerning an individual or similarly significantly affects an individual shall make readily available meaningful information regarding: (a) the fact that an AI system is being used in a decision-making process; (b) the intended purpose(s); (c) the types of data sets that are used and generated by the AI system; and (d) whether and to what extent the decision-making process may include human participation.

2.2 The information set forth in Section 2.1 should be made readily available to the affected individual before such automated decision-making process occurs in order to provide the individual with an opportunity to assess whether or not to seek a human-centric alternative decision-making process.

3 Explainability

3.1 Organisations that make available or use an AI system in decision-making processes which produce legal effects concerning an individual or similarly significantly affects an individual shall make readily available to such individuals information in objectively clear terms that explains how a decision/outcome was reached, with, at a minimum: a) the information set forth in Section 2.1 above; b) information that offers meaningful interpretability of the algorithmic logic of the AI system; c) meaningful information to understand the decision/outcome; and d) information regarding how the individual may contest the decision or outcome.

3.2 The information set forth in Section 3.1 should be made readily available to an affected individual promptly after such automated decision-making process occurs in order to provide the affected individual with an opportunity to assess whether or not to challenge the decision or outcome.

4 Gradual and contextual approach

4.1 The intensity of the transparency and explainability obligations will depend on a variety of factors, including the nature of the data involved, lack of human participation in the decision-making, the result of the decision and its consequences for the affected individual.

4.2 Ultimately, transparency and explainability must balance the rights, interests and reasonable expectations of the person subject to the decision with the legitimate interests of the organisation making the decision and considerations of overall societal benefit.

4.3 The intensity of the transparency and explainability obligations will generally be higher where the AI system is made available or used in relation to lay persons who are unlikely to understand the technology rather than with an expert whose understanding of the system may be more easily established. Moreover, the intensity of the transparency and explainability obligations will generally be higher where an AI system is used by a public sector organization in the context of enforcing legal obligations rather than by a private sector organisation in the context of offering services.

4.4 The intensity of the transparency and explainability obligations will generally be higher where sensitive personal data is used or where the outcome of the decision will have a material impact on the affected individual's legal or human rights or similarly significantly affects an individual. The intensity of these obligations will generally be lower where non-sensitive personal data or de-personalised data is used or where the impacts on the affected individual's legal or human rights are relatively inconsequential.

4.5 In situations giving rise to high intensity transparency and explainability obligations, organisations that make available or use an AI system in decision-making processes affecting individual rights should, in addition to the information set forth in Sections 2.1 and 3.1 above, make readily available to such individuals meaningful information regarding: a) the traceability and auditability of the algorithmic logic of the AI system, and b) the testing methods used to promote the principles within this policy framework.

5 Transparency and explainability by design

5.1 Organisations that develop AI systems should ensure that the system architecture, algorithmic logic, data sets, testing methods, and all related development and operational policies and procedures serve to incorporate and embed transparency and explainability by design in accordance with national laws and consistent with relevant industry standards. In so far as is reasonably practicable, such systems should aim to be designed from the outset and maintained to promote meaningful transparency and explainability that complements the intended purpose(s) of the AI system.

4.2 The design and development methodologies adopted in Section 5.1 should have the flexibility to embrace evolving industry standards, providing ongoing iterative improvements in transparency and explainability in parallel with advancement in the state of the art during the lifecycle of the AI system.

4.3 Since embedding transparency and explainability into AI system design requires extensive planning and multi-disciplinary expertise, organisations should develop frameworks to assist programmers and developers to design and develop AI systems that possess the desired values and to help reconcile the tensions that exist between accuracy, cost and explainability.

6 Technological neutrality

6.1 The use of an AI system by an organisation does not increase or reduce the procedural and substantive requirements that would otherwise apply if the decision-making process were controlled by a human.

Endnotes

1 See https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

2 For example, some insurers now use AI systems to screen applicants, propose and price policies, and submit and process claims. Even in cases where an insurer did not originally or entirely develop the AI system in-house, it should remain responsible for its use, based on its own governance model and internal policies, procedures, and controls. See The Geneva Association, "Promoting Responsible Artificial Intelligence in Insurance," https://www.genevaassociation.org/sites/default/files/research-topics-document-type/pdf_public/ai_in_insurance_web_0.pdf.

3 This reference to industry standards means frameworks that are created collaboratively by standards bodies and made available for public use, not best practices that are recognized over time with the benefit of hindsight. See, e.g. PCI Security Standards Council, https://www.pcisecuritystandards.org/pci_security/standards_overview (for payment card data security standards).

4 See https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf. This report was issued to advance public discussion on AI and facilitate the creation of IEEE standards.

5 See https://www.nist.gov/document/four-principles-explainable-artificial-intelligence-nistir-8312.

6 See https://www.nist.gov/document/report-plan-federal-engagement-developing-technical-standards-and-related-tools.

7 See https://www.darpa.mil/program/explainable-artificial-intelligence.

8 See Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning, Ruth M.J. Byrne, available at https://www.ijcai.org/Proceedings/2019/876.

9 It may be possible to disclose some general information relevant to explainability before a decision is made, such as high-level factors that will be taken into consideration in all circumstances. See, e.g. https://www.google.com/search/howsearchworks/algorithms/ (for an overview of how Google's search algorithm works). Yet such information may not provide a sufficient basis to enable individuals to protect or enforce their rights.

10 This is an area ripe for industry standards to establish a technological baseline that evolves over time.

11 See https://www.ipc.on.ca/wp-content/uploads/resources/7foundationalprinciples.pdf (identifying the seven foundational principles of privacy by design); and GDPR Article 25 (requiring data protection by design and by default).

12 Consistent with the gradual, context approach in Section 6, we recognize that not all AI systems may need to offer explanations due to the limited nature of their use. Similarly, we appreciate that some AI systems may simply be too complex to be explainable (at least not without a separate, cloned AI system that interprets insights from, and offers explanations for, the proverbial black box). Despite these realities, we believe it prudent and appropriate to anchor AI to a default of explainability by design, especially looking forward as new AI systems are developed.

13 https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai.

Principle 4

FAIRNESS AND NON-DISCRIMINATION

Responsible AI

A GLOBAL POLICY FRAMEWORK

Principle 4 Commentary

FAIRNESS AND NON-DISCRIMINATION

CHAPTER LEAD
Diego Fernández | Marval O'Farrell Mairal, Argentina

Jason Haas | Ervin Cohen & Jessup LLP, USA

Alexander Tribess | Weitnauer Partnerschaft mbB, Germany

Patricia Shaw | Beyond Reach, UK

Salvatore Orlando | Ughi e Nunziante Studio Legale, Italy

Inés O'Farrell | LLM candidate at Harvard Law School

Introduction

The standards set forth in *Responsible AI: A Global Policy Framework* with respect to fairness and non-discrimination continue to offer important guidance to governments and organisations considering how to best address the issues of unwanted bias in AI systems. Yet a number of publications over the last year have focused attention on certain aspects of the existing principles that require some refinement or expansion. The revised principles seek to address these aspects while remaining true to the original objectives for Principle 4. The primary challenge moving forward in this area, however, is not to define objectives that are already widely shared. Rather, it is developing the approaches and methods that organisations and governments can use to minimise the existence of bias in the use of AI and to seize the possibilities that AI systems offer to achieve more fair and equal outcomes across society.

Global developments on bias and discrimination

AI HLEG ethics guidelines for trustworthy AI

In June 2018, the European Commission established an independent high-level expert group (the AI HLEG) that released influential ethical guidelines on 8 April 2019.[1]

The Guidelines start with the premise that there are fundamental rights that should be considered in the development, deployment and use of AI systems, one of which is "equality, non-discrimination and solidarity." These rights are broader than simply nondiscrimination and seek both to prevent unfairly biased outputs and to ensure "adequate respect" for the rights of potentially vulnerable groups such as women, persons with disabilities and consumers.

The Guidelines further require fairness in the development, deployment and use of AI systems. They explain that this has two dimensions: substantive and procedural. The substantive dimension seeks to

ensure fair and unbiased outcomes and to allocate the burdens and benefits of AI in a just and equal manner. The Guidelines recognize that "[i]f unfair biases can be avoided, AI systems could even increase societal fairness." The procedural dimension of fairness, in the view of the AI HLEG, involves the ability of an individual affected by a biased AI system to challenge a potentially biased decision and to seek a remedy for any harm that has been suffered. Two key requirements for such fairness is an ability to identify who is accountable for the AI's decisions and for the decision-making process by the AI to be explainable.

The Guidelines discuss how the various ethical requirements, including diversity, non-discrimination and fairness, support each other and emphasise that they need to be "continuously evaluated and addressed throughout the AI system's lifecycle." If possible, identifiable bias should be removed from training data when it is initially collected, but having the proper process in place to oversee the development of an AI can help to prevent bias from arising later in the design process. The AI HLEG also calls not only for the interests of all affected stakeholders to be considered during the design phase, but for those stakeholders actually to be involved throughout the life cycle of an AI system. The AI HLEG further encourages the hiring of persons of diverse backgrounds, cultures, and disciplines in AI development to ensure a range of perspectives and opinions. The Guidelines include a draft list of questions for organisations developing AI to consider with respect to fairness and non-discrimination issues. The EU White Paper on Artificial Intelligence, published in February 2020 (and mentioned below) confirmed that The HLEG is in the process of revising its guidelines in light of public feedback and consultation on this list of questions and (as at the date of publication of this update) is due to finalise its work imminently.

Hambach Declaration on Artificial Intelligence

This April 3, 2019 white paper by the Independent Federal and State Data Protection Supervisory Authorities of Germany stresses that AI systems are highly dependent on training data, and that discriminating effects can result from flawed data or programming.[2] Yet it recognises that discriminatory effects may not always be readily apparent in the design stage. It recommends both "an assessment of risks for the rights and freedoms of people" before an AI system is first made available that would utilise reliable techniques to identify any concealed bias, and ongoing monitoring during the use of the AI system to detect any risk of biased outcomes.

Chinese governance principles for a new generation of AI

On 17 June 2019, the National New Generation AI Governance Expert Committee established by China's Ministry of Science and Technology issued proposed principles for AI governance and "responsible AI."[3] One of the eight principles was "Fairness and Justice." That principle set a goal for Chinese developers to improve the technology and management methods incorporated in AI systems to "eliminate bias and discrimination in the process of data acquisition, algorithm design, technology development, product R&D, and application."

White House OMB draft memo on the regulation of AI

On 7 January 2020, the U.S. White House Office of Management and Budget released a draft memorandum[4] to offer U.S. agencies guidance regarding the development of both regulatory and non-regulatory approaches towards systems and businesses that use AI. The Memo encourages the adoption of polices that both promote progress in AI technology and innovation and protect American national interests and values.

In general, the memo advocates an approach that promotes AI technology to improve outcomes compared to existing processes. It cautions that agencies should not hold AI to "impossibly high standards." Instead, they are advised to consider carefully "the full societal costs, benefits, and distributional effects" of employing AI as compared to existing systems and procedures. Agencies are also counselled to consider whether the types of errors generated by the AI would differ in nature from the existing systems and how the level of risk arising under the AI would compare to the degree of risks that were accepted under current procedures.

The memo extends this approach to the assessment of potentially biased outcomes from AI systems. It highlights the potential for AI to improve outcomes by "reducing present-day discrimination caused by human subjectivity." Yet it recognises that biased AI systems can also result in discriminatory outcomes. It thus encourages U.S. agencies to consider both the fairness and discriminatory impact of a given type of AI system as well as whether the system might reduce current levels of discrimination.

UK ICO guidance on the AI auditing framework

On 14 February 2020, the UK Information Commissioner's Office released draft guidance that includes auditing tools and procedures for compliance professionals and technical specialists to evaluate the performance of AI systems, including with respect to bias and non-discrimination.[5] It discusses the different types of statistical errors that can generate bias in AI and offers several analytical approaches that can be used to detect bias in the operation of AI systems. The Guidance emphasises the need for organisations to document an approach to bias and discrimination mitigation even before beginning the design phase for a new AI. It advocates robust testing of anti-discrimination measures to monitor the performance of the system on an ongoing basis. It also recommends setting acceptable tolerance levels against which to measure the AI system's performance that, if exceeded, will trigger remedial steps and even terminate use of the system if necessary. The ICO observed that, with the proper steps and safeguards, AI systems can be used to identify and correct discrimination that existed before the AI was deployed.

The EU White Paper on Artificial Intelligence

On 19 February 2020, the European Commission published its White Paper on Artificial Intelligence,[6] which presents the fullest statement to date of the Commission's approach to the regulation of artificial intelligence. That paper recognised that the improper use of AI could result in violations of "fundamental rights," including the right to "non-discrimination based on sex, racial or ethnic origin, religion or belief, disability, age or sexual orientation." It acknowledged that bias and discrimination are risks in any activity but notes that a biased AI could have a much larger impact. The White Paper also highlights the risk that

bias could be introduced into an AI system after implementation if the AI system "learns" the wrong lessons from the patterns in the dataset it utilises to make decisions.

Summary and conclusions

Revisions to Principle 4

The revisions to Principle 4 offer a simultaneously more hopeful and more realistic view of the challenge posed in addressing unfair bias in AI, whether such bias simply builds on existing patterns of societal discrimination or presents new problems that arise from the way in which an AI system was developed, made available or used. On one hand, it is not enough for AI systems to simply avoid making outcomes worse than they had been in a pre-AI world. There are real prospects for AI systems to make significant progress in the areas of fairness and non-discrimination in many spheres of life and society. On the other hand, unwanted bias can be introduced into an AI system in many ways. As recent publications such as the HLEG Ethics Guidelines for Trustworthy AI have emphasised, it requires concerted and ongoing attention throughout the life cycle of an AI system to avoid bias being introduced by core algorithms, to detect any bias in procedures or outcomes that survived the training and design phase or which may arise subsequently during the operation of an AI, and to remedy the negative consequences suffered by individuals whose lives have been impacted by the biased and unfair decision-making of a flawed AI system.

Progress in these areas may be incremental, but it is reasonable to expect that new AI systems should be systematically assessed before deployment for the risk of bias, and rigorously compared to existing procedures and outcomes to ensure that they do not exacerbate existing problems with bias and increase the unfair treatment of certain groups.

Many of these refinements, and particularly the new emphasis on having an ongoing process to detect and correct bias at every stage of the development and use of an AI system, are consistent with the conclusions of a number of the recent publications discussed above.

Principle 4
Fairness and Non-Discrimination

Organisations that develop, make available or use AI systems and any national laws that regulate such use shall ensure the non-discrimination of AI outcomes, and shall promote appropriate and effective measures to safeguard fairness in AI use.

1 Awareness and education

1.1 Awareness and education on the possibilities and limits of AI systems is a prerequisite to achieving fairer outcomes.

1.2 Organisations that develop, make available or use AI systems should take steps to ensure that users are aware that AI systems reflect the goals, knowledge and experience of their creators, as well as the limitations of the data sets that are used to train them.

2 Technology and fairness

2.1 Carefully designed AI systems offer the possibility of more consistently fair and non-discriminatory outcomes than are achievable in systems that rely on human decision-making.

2.2 Decisions based on AI systems should be fair and non-discriminatory, judged against the same standards as decision-making processes conducted entirely by humans.

2.3 The use of AI systems by organisations that develop, make available or use AI systems and Governments should not serve to exempt or attenuate the need for fairness, although it may mean refocusing applicable concepts, standards and rules to accommodate AI.

2.4 Users of AI systems and persons subject to their decisions must have an effective way to seek remedy in discriminatory or unfair situations generated by biased or erroneous AI systems, whether used by organisations that develop, make available or use AI systems or governments, and to obtain redress for any harm. Taking into consideration the societal impacts of unfair AI, collective remedies could be a useful tool to address bias or unfairness.

3 Development and monitoring of AI systems

3.1 AI development should be designed to prioritise fairness and non-discrimination. This would involve addressing algorithms and data bias from an early stage and continuously throughout the entire lifecycle of the AI system with a view to ensuring fairness and non-discrimination.

3.2. Before making available or using an AI system, organisations should systematically assess the expected performance of the AI system with respect to potentially unlawful or unfair discrimination as compared to the performance of the processes currently in use.

3.3. Organisations that develop, make available or use AI systems should remain vigilant to the dangers posed by bias. This could be achieved by establishing ethics boards and codes of conduct, and by adopting industry-wide standards and internationally recognised quality seals.

3.4. AI systems with an important social impact could require independent reviewing and testing on a periodic basis.

3.5. In the development and monitoring of AI systems, particular attention should be paid to disadvantaged groups which may be inadequately or unfairly represented in the training data.

4 A comprehensive approach to fairness

4.1 AI systems can perpetuate and exacerbate bias, and have a broad social and economic impact in society. Addressing non-discrimination and fairness in AI use requires a holistic approach. In particular, it requires:

- **i.** the close engagement of technical experts from AI-related fields with statisticians and researchers from the social sciences; and
- **ii.** a combined engagement between governments, organisations that develop, make available or use AI systems and the public at large.

4.2 The Fairness and Non-Discrimination Principle is supported by the Transparency and Accountability Principles. Effective fairness in use of AI systems requires the implementation of measures in connection with both these Principles.

Endnotes

1 European Commission, High-Level Expert Group on Artificial Intelligence "Ethics Guidelines for Trustworthy AI" (8 April 2019), COM (2019), online: https://ec.europa.eu/futurium/en/ai-alliance-consultation.

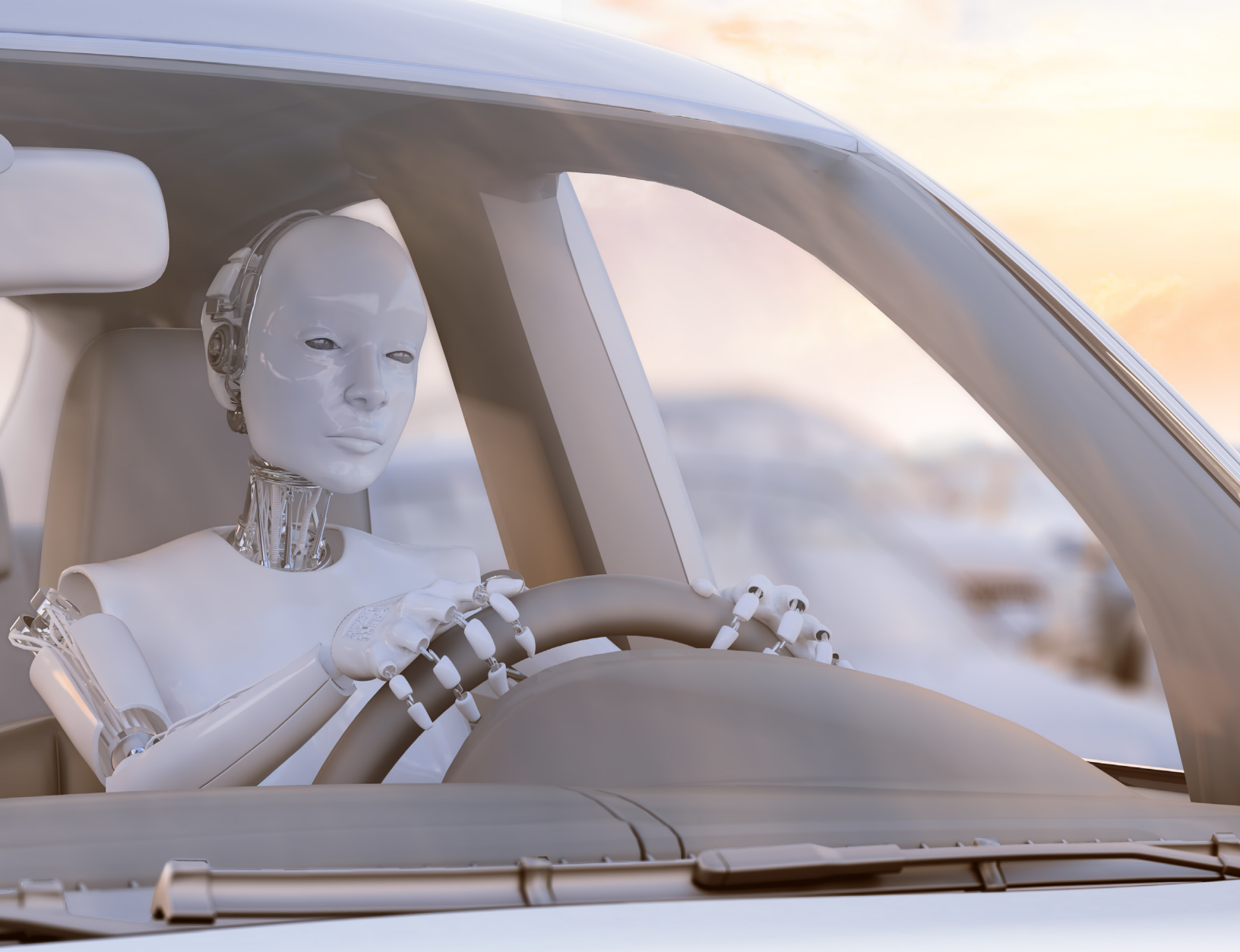
2 Resolution of the 97th Conference of the Independent Federal and State Data Protection Supervisory Authorities of Germany, Hambach Castle, "Hambach Declaration on Artificial Intelligence" (3 April 2019), online https://www.datenschutz-berlin.de/fileadmin/user_upload/pdf/publikationen/DSK/2019/2019-DSK-Hambach_Declaration_AI-en.pdf.

3 China Ministry of Science and Technology, National New Generation AI Governance Expert Committee, "Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence, (7 June 2019), online: https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-expert-group-offers-governance-principles-responsible-ai/.

4 White House Office of Management and Budget (OMB) and the Office of Science and Technology Policy (OSTP), Vought, Russell. T., "Draft Memorandum re: Guidance for Regulation of Artificial Intelligence Applications," (19 October 2019), online: https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf.

5 UK Information Commissioner's Office, "Guidance on the AI auditing framework: Draft guidance for consultation" (14 Feb. 2020), online: https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf.

6 European Commission, "White Paper on Artificial Intelligence" (19 February 2020), COM (2020) 65 final, online: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.



Principle 5

SAFETY AND RELIABILITY

Responsible AI

A GLOBAL POLICY FRAMEWORK

Principle 5 Commentary

SAFETY AND RELIABILITY

Christian Frank | Taylor Wessing, Germany
Louis Jonker | Van Doorne, Netherlands
Stuart P. Meyer | Fenwick & West LLP, United States

Introduction

Since we finalised our work on the first edition, very many contributions, reports and guides have been published further advancing the public discussion and understanding of the safety and reliability aspects of artificial intelligence.[1] We have gained new insights, based on which we have supplemented the principles on Safety and Reliability to make individual aspects more precise and more clearer.

Overview on the key developments since original publication

On April 4, 2019, the EU Parliament's Panel for the Future of Science and Technology (STOA) published its report "A governance framework for algorithmic accountability and transparency" specifying inter alia policy options for the governance of algorithmic transparency and accountability, based on an analysis of the social, technical and regulatory challenges posed by algorithmic systems.[2]

On April 8, 2019, the EU Commissions High-Level Expert Group (HLEG) on AI presented Ethics Guidelines for Trustworthy Artificial Intelligence.[3] This followed the publication of the guidelines' first draft in December 2018 on which more than 500 comments were received through an open consultation. According to these Guidelines, trustworthy AI should inter alia be robust—both from a technical perspective while taking into account its social environment. The HLEG Ethics Guidelines lay out three components for trustworthiness, the third of which is relating to robustness, both from a technical and social perspective. Such technical robustness and safety includes "resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility." It also refers to a pilot version of an assessment list specifying aspects in relation to all these criteria, which practitioners may use to improve the trustworthiness of any respective system.[4]

In the following month, on May 28 the Beijing Academy of Artificial Intelligence (BAAI) released the "Beijing AI Principles," for the "research, development, use, governance, and long-term planning of AI."[5] Amongst other things, these principles promote safety and reliability by referring to *"continuous efforts…. to improve the maturity, robustness, reliability, and controllability of AI systems, so as to ensure the security for the data, the safety and security for the AI system itself, and the safety for the external environment where the AI system deploys."* This was followed by a draft "joint pledge" (公约) on self-discipline in the artificial intelligence industry by China's AI Industry Alliance (AIIA).[6] In Article 5 of the joint pledge, the signatories commit to ensuring that AI systems operate securely/safely, reliably, and controllably through-

out their lifecycle. Furthermore, they undertake to evaluate system security/safety and potential risks, and continuously improve system maturity, robustness, and anti-tampering capabilities and to ensure that the system can be supervised and promptly taken over by humans to avoid the negative effects of loss of system control. Furthermore, they aim to continuously improve the transparency of AI systems which including a reminder on the explainability, predictability, traceability, and verifiability of system decision-making processes.

On June 10, 2019, the UK government published a guide to using artificial intelligence in the public sector.[7] The initiative was led by the Office for Artificial Intelligence (OAI) and the Government Digital Service (GDS), with The Alan Turing Institute's public policy programme contributing guidance on AI ethics and safety. It covers in particular, how the public sector can best use AI and explains how the potential uses for AI in the public sector are significant, but must be balanced with ethical, fairness and safety considerations. The guide identifies the potential harms caused by AI systems and proposes concrete, operationalisable measures to counteract them. It stresses that public sector organisations can anticipate and prevent these potential harms by stewarding a culture of responsible innovation and by putting in place governance processes that support the design and implementation of ethical, fair, and safe AI systems

On June 21, 2019, Executive Office of US President Trump published the 2019 Update to the US Artificial Intelligence Research and Development Strategic Plan released in 2016.[8] In relation to its strategy 4 on the Safety and Security of AI Systems, it is further emphasised that *"the notion of 'safety (or security) by design' might be misleading to the extent that these are only concerns of system designers: instead, they must be considered throughout the system lifecycle, not just at the design stage, and so must be an important part of the AI R&D portfolio."*

On July 2, 2019, a team of the Universities of Bonn and Cologne and Fraunhofer IAIS presented its interdisciplinary approach in a white paper aiming to form a basis for the further development of AI certification.[9] In an interdisciplinary approach, the authors explain the identified fields of action from a philosophical, ethical, legal and technological point of view emphasising in their chapter on legal requirements the aspects responsibility, traceability, and liability for AI applications. The paper serves as a contribution to the social debate on the trustworthy use of AI and the further development of certification.

On January 27, 2020, the German Association of Technical Inspection Agencies published a survey on consumers' expectations in relation to security and Artificial Intelligence pursuant to which a large majority of the surveyed are said to be of the opinion that AI products should only be brought onto the market once their safety has been verified by independent bodies and that the state should adopt laws and regulations to regulate AI.

On 19 February 2020, the European Commission published its White Paper aiming to foster and European ecosystem of excellence and trust in artificial intelligence, and more specifically, a Report on safety and liability implications of AI, the Internet of Things and Robotics.[10] The report aims to identify and examine the broader implications for and potential gaps in the liability and safety frameworks and to facilitate the discussion and a part of the broader consultation of stakeholders. In its conclusion, it is noted that the current European product safety legislation contains a number of gaps to be addressed, in particular in the General Product Safety Directive, Machinery Directive, the Radio Equipment Directive and the New

Legislative Framework. The new challenges in terms of safety are said to create also new challenges in terms of liability, which need to be fixed to ensure the same level of protection compared to victims of traditional technologies, while maintaining the balance with the needs of technological innovation. This shall help create trust in these new emerging digital technologies and create investment stability.

On March 24, 2020, the South African Policy Action Network, an initiative of the Human Sciences Research Council (HSRC), supported by the South African Department of Science and Innovation published several topical Guides on AI & Data.[11] In its Guide No 5 on AI, Biometrics and Securitisation in Migration Management it is in particular recommended requiring the users of biometrics and AI systems to determine the technical reliability thereof in generating data for decision making.

The German Association for Electrical Engineering, Electronics and Information Technology and the Bertelsmann Stiftung published a framework to operationalise AI ethics on April 4, 2020.[12] The interdisciplinary team authoring the report are inter alia emphasizing reliability as precondition for trust, predictability and safety aspects as robustness and resilience and cybersecurity as confidentiality, integrity and availability.[13]

Summary and conclusions

Some of the initiatives taken are either very sector specific or only non-binding promises. In view of the unstoppable globalization and the impact AI systems may have on the societies as a whole and the individual, it seems preferable to aim for wider reaching binding standards. International private bodies such as International Organization for Standardization have already published and continue to work on respective standards from a technical point of view. As they lack regulatory authority, their publications may gain relevance as de facto standards or get implemented via incorporation in respective agreements. However, such assertion requires more time, so public regulation seems more appropriate to ensure in particular sufficient public safety.

Amendments to Principle 5

Changes to paragraph 3.1

Data is the essential fuel of every AI system. Secure and reliable functioning can be affected if a system that is *per se* capable to operate properly is fed, trained and run with datasets which are not correct, representative and generalisable. So we specified paragraph 3.1 S. 1 accordingly.

AI systems are designed to obtain decision-making power and autonomously make decisions. The highest engineering efforts and skills will probably not be sufficient to rule out that bad decisions will happen. Progress always depends on learning from mistakes and avoiding them in the future. Transparency requirements will speed up and facilitate error analysis enabling an improvement of the safety and reliability of the AI systems in use. Apart from the obvious technical side of transparency, it reflects a traditional legal concept: Whenever lawyers are unable to specify a future result but desire the probability of a proper outcome, they tend to specify the nature and manner of the approach: This is the legal concept behind,

e.g. “due process” requirements in criminal law, “clinical testing” requirements of drug approvals, or the typical way research and development agreements are designed. Instituting and following a proper procedure is further relevant when deciding whether a regrettable outcome has indeed been unforeseeable so that the responsibility for them may be limited. Hence, we emphasised this aspect in adding a separate sentence to paragraph 3.1.

Changes to paragraph 4.2

The reason for including paragraph 4.2 is based on a related concern. Continuous improvement will increase the safety and reliability for which obligations on continuing monitoring are essential. They will build overall trust that the approach is guided by a sense of responsibility being important for society to accept the delegation of decision making powers to AI based systems.

Changes to paragraph 4.3

Finally, the amendment to paragraph 4.3 specifies the recommendation of maintaining product safety by calling for duties of continuous monitoring with human oversight.

Principle 5
Safety and Reliability

Organisations that develop, make available or use AI systems and any national laws that regulate such use shall adopt design regimes and standards ensuring high safety and reliability of AI systems on one hand while limiting the exposure of developers and deployers on the other hand.

1 Require and/or define explicit ethical and moral principles underpinning the AI system

1.1 Governments and organisations developing, making available or using AI systems should define the relevant set of ethical and moral principles underpinning the AI system to be developed, deployed or used taking into account all relevant circumstances. A system designed to autonomously make decisions will only be acceptable if it operates on the basis of clearly defined principles and within boundaries limiting its decision-making powers.

1.2 Governments and organisations developing, making available or using AI systems should validate the underpinning ethical and moral principles as defined periodically to ensure ongoing accurateness.

2 Standardisation of behaviour

2.1 Governments and organisations developing, making available or using AI systems should recall that ethical and moral principles are not globally uniform but may be impacted e.g. by geographical, religious or social considerations and traditions. To be accepted, AI systems might have to be adjustable in order to meet the local standards in which they will be used.

2.2 Consider whether all possible occurrences should be pre-decided in a way to ensure the consistent behaviour of the AI system, the impact of this on the aggregation of consequences and the moral appropriateness of "weighing the unweighable" such as life vs. life.

3 Ensuring safety, reliability and trust

3.1 Governments should require and organisations should test AI systems thoroughly to ensure that they reliably and robustly adhere, in operation, to the underpinning ethical and moral principles and have been trained with data which are curated and are as 'error-free', 'bias-free' as practicable, given the circumstances. This includes requirements on procedural transparency and technical transparency of the development process of the AI system and the data uses in that respect, as well as the explainability of the decision-making process an AI system will apply when in operation.

3.2 Governments are encouraged to adjust regulatory regimes and/or promote industry self-regulatory regimes for allowing market-entry of AI systems in order to reasonably reflect the positive exposure that may result from the public operation of such AI systems. Special regimes for intermediary and limited admissions to enable testing and refining of the operation of the AI system can help to expedite the completion of the AI system and improve its safety and reliability.

3.3 In order to ensure and maintain public trust in final human control, governments should consider implementing rules that ensure com-

prehensive and transparent investigation of such adverse and unanticipated outcomes of AI systems that have occurred through their usage, in particular if these outcomes have lethal or injurious consequences for the humans using such systems. Such investigations should be used for considering adjusting the regulatory framework for AI systems; in particular to develop a more rounded understanding of how such systems should gracefully handover to their human operators.

4 Facilitating technological progress at reasonable risks

4.1 Governments are encouraged to consider whether existing legal frameworks such as product liability require adjustment in light of the unique characteristics of AI systems.

4.2 As AI systems might be partially autonomous, organisations developing, deploying or using such systems should pursue continuous monitoring of systems deployed and/or used, allowing human operators to interrupt unanticipated alterations.

4.3 Governments should support and participate in international co-ordination (through bodies such as the International Organisation for Standardisation (ISO) and the International Electrotechnical Commission (IEC)) to develop international standards for the development and deployment of safe and reliable AI systems. Governments are further encouraged to contemplate requirements on continuous monitoring with human oversight as part of their regime balancing encouragement of progress vs. risk avoidance.

Endnotes

1 See, e.g. the overview Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar, *"Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI."* Berkman Klein Center for Internet & Society, 2020 http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420; and the summary below.

2 See https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf.

3 See https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

4 See https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines.

5 See https://www.baai.ac.cn/news/beijing-ai-principles-en.html.

6 See the Chinese language original at https://mp.weixin.qq.com/s/x7HTx4AR6oNBWwWxUpnSuQ and the English translation at https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-ai-alliance-drafts-self-discipline-joint-pledge/.

7 See https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector.

8 See https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf.

9 See https://www.iais.fraunhofer.de/content/dam/iais/KINRW/Whitepaper_Thrustworthy_AI.pdf.

10 COM(2020) 64 final, see https://ec.europa.eu/info/sites/info/files/report-safety-liability-artificial-intelligence-feb2020_en_1.pdf.

11 See https://policyaction.org.za/ai-data-topical-guide-series.

12 See https://www.bertelsmann-stiftung.de/de/publikationen/publikation/did/from-principles-to-practice-wie-wir-ki-ethik-messbar-machen-koennen.

13 See page 29.

Responsible AI

A GLOBAL POLICY FRAMEWORK

Principle 6

OPEN DATA AND FAIR COMPETITION

Principle 6 Commentary

OPEN DATA AND FAIR COMPETITION

CHAPTER LEAD
John Buyers | Osborne Clarke LLP, UK

Nicola Benz | Froriep, Switzerland

Per-Kaare Svendsen | Kvale, Norway

Dr Sonja Dürager | bpv Hügel Rechtsanwälte, Austria

Cornelia Mattig | Froriep, Switzerland

Introduction

Since the publication of *Responsible AI: A Global Policy Framework*, there have been a number of developments in the open data space which have necessitated perhaps a fuller update than other sections of this revised publication. Firstly, it has become clear that principles similar to the ones we advocated for in respect of Open Data have become a key part of the EU's new legislative strategy on Artificial Intelligence, published at the beginning of 2020. Secondly, as the sharing of data between enterprises has become more commonplace (rising hand in hand with the increase in adoption of AI based solutions) we have focussed on some of the commercial strategies to achieve this, in the absence of any consistent global legislative approach. It is evident that we are at an inflection point in terms of the way in which data (and data sets) are shared between businesses (and indeed other entities) and the concept of a trusted data intermediary or "data trust" is likely to play more of a part in these exchanges in the future. It is, as yet, too early to say quite how such data trusts will be shaped and constituted, but we raise a flag at this stage to say that such concepts are likely to dramatically change the way in which we exchange and commoditise data within our global connected marketplace.

Revisions to Principle 6

We were conscious of feedback from our clients and peers on the last version of the principle which perhaps took too much of a sweeping view of open data and neglected the rather delicate interface between this, the need for commercial enterprise to maintain ownership in intellectual property terms (as well as the value of invested effort) and individual privacy when dealing with personal data. Nevertheless, it is clear that data portability is becoming a significant focus for governments and regulators, and the private sector—especially so in light of the recent COVID-19 pandemic.

We have consequently adjusted the principle to more accurately reflect this renewed focus on portability and also to align it more closely to the proposed approaches advocated by the EU in its recent AI white paper and Microsoft in its Open Data Campaign.

What the substances of the changes actually reflect therefore is a call to organisations and enabling national policy frameworks to facilitate portability and open access to data sets and corresponding AI systems, especially in circumstances where such datasets or systems are "significant and important" or which would advance the state of the art. This is clearly advocated to be made subject to intellectual property rules and with respect to the levels of investment and ownership that have been made in such assets.

With the above in mind, we now move to a discussion of recent major developments.

The EU White Paper on Artificial Intelligence

On 19 February 2020, the European Commission published its White Paper on Artificial Intelligence[1] and its European Strategy for Data,[2] both of which place a strong emphasis on open access to data.

The EU's White Paper on AI is the fullest statement to date the European Commission's approach to the regulation of artificial intelligence—an overriding legislative priority kicked off by Commission President Ursula Van Der Leyen at the start of her term. Much in the same way as the GDPR, the approach it advocates will be highly influential and is likely to lead to global changes in the way in which AI systems (and their underlying datasets) are utilised.

The White Paper promotes open access to data as one of the important drivers for developing artificial intelligence technology. It rightly notes that without data the development of AI and other digital applications is not possible and there is an opportunity for Europe to be at the forefront of AI transformation by positioning itself well in relation to the vast amounts of data yet to be generated.

The focus on data in the White Paper is clearly stated to be in tandem with and not at the expense of fundamental European values such as data protection. The White Paper also promotes responsible data management practices and compliance of data with "FAIR" principles (Findable, Accessible, Interoperable and Re-usable) in order to build trust and ensure data re-usability.

Furthermore, the White Paper announces under the heading "Securing Access to Data and Computing Infrastructure" that more than EUR 4 billion of funding will be available under the Digital Europe Programme to support data and cloud infrastructure, as well as high-performance and quantum computing, including edge computing and AI.

With this dual approach of promoting open access to data in compliance with FAIR principles and funding for (open access) data infrastructure, the White Paper takes a very supportive stance on open data, while not prescribing it. For the private and scientific sectors, participation in open data access projects is encouraged, principally through the availability of funding, but is not mandatory. For the public sector, starting with the European Commission itself, there are more ambitious plans for open access to data, as had already been set out in the European Commission Digital Strategy[3] published in 2018, which embodies a vision for the Commission to become a digitally transformed, user-focused and data-driven administration by 2022.

European Strategy for Data

The European Strategy for Data focuses on how the EU can acquire a leading role in the data economy by creating a single European data space. This is expressed as a single market for data, open to data from across the world—where personal as well as non-personal data, including sensitive business data, are secure. The overriding approach is to ensure that the public sector and businesses have easy access to high-quality data, boosting growth and creating value, while minimising carbon emissions and environmental footprints.

A distinction is made in the strategy between government to business data sharing, business to business data sharing, business to government data sharing and data sharing between public authorities, each category bringing its own challenges.

Of particular interest are the hurdles identified for business to business data sharing, and by analogy also business to government data sharing and the possible approaches to promote such sharing, namely: a lack of economic incentives (including the fear of losing a competitive edge), lack of trust between economic operators that the data will be used in line with contractual agreements, imbalances in negotiating power, the fear of misappropriation of the data by third parties, and a lack of legal clarity on who can do what with the data (for example for co-created data, in particular IoT data).

An Expert Group[4] created by the Commission has recommended approaches to promote business-to-government data sharing including the creation of national structures for B2G data sharing and the development of appropriate incentives to create a data-sharing culture as well as, perhaps more controversially, a suggestion to explore an EU regulatory framework to govern the public sector's re-use for the public interest of privately-held data.

With regard to government-to-business data sharing, the Directive on Open Data and the Re-use of Public Sector Information[5] aims to reduce market entry barriers, in particular for SMEs. It does this by generally capping the amount that public bodies can charge for the re-use of their data to no more than the marginal costs of dissemination. The Directive must be implemented into Member State laws by 17 July 2021.

UK Smart Data Review

On 11th June 2019, the UK Government released a consultation paper advocating new requirements for service providers to share consumer data. This paper is part of the UK's digital strategic focus and recommends that data in regulated markets, such as financial services, energy and telecoms be made more easily accessible to third parties to enable consumer choice. Data portability is considered key to this strategy, and the paper considers that models such as the Open Banking initiative should be adapted to facilitate the transfer of consumer data between services in a commonly used machine-readable format.[6]

Developments in the United States—Microsoft's Open Data Campaign

In the US, Open Data initiatives have been largely private sector led. Microsoft recently launched its Open Data Campaign[7] at the end of April 2020. Microsoft's stated ambition is to address the growing "data divide" between businesses that have access to data and those that do not, and to improve global collaboration between data holders. Microsoft is working closely with the UK's Open Data Institute and The Governance Lab at New York University Tandon School of Engineering to promote this campaign. Microsoft itself has stated that it will lead by example and has published a set of five principles to guide what it terms "trusted data collaboration"[8]

Commercial data sharing agreements (DSA)

Since the publication of *Responsible AI: A Policy Framework*, and as we note in the introduction to this update, the evolution of commercial data sharing has continued apace.

As we referred to in the original version of *Responsible AI*, Chapter (para 6 II D), there have been a number of calls for legal frameworks to facilitate sharing of data, including efforts to create common standards for sharing open data and model agreements to facilitate use of large datasets for computational purposes and development of AI.

Data comprise a vast span of different categories each with their own characteristics in terms of data sources, legal regulation and commercial applicability. The natural instincts of commercial enterprise to keep data proprietorial and closed is seen increasingly as an impediment to competition and trade. Despite this data sharing arrangements are now found in all aspects of the digital economy, both in traditional vertical value chains (i.e. between a supplier of advanced services based on machine learning and AI and its customers) or in horizontal agreements between two parties collaborating to develop new technology or services.

As data sharing arrangements have expanded beyond traditional boundaries to encompass the sharing of more and more data, such commercial arrangements have become more fluid and relationships need to be managed and regulated between parties that are not interacting directly. This has led to the creation of collaboration platforms, some with scientific and academic focus and others with commercial focus within one or more traditional industries.

In the absence of legal external restraints for sharing data (subject to copyright, data protection laws and trade secrets laws as mentioned in section 6 II A of our original chapter) access to—and the use of data may nevertheless be subject to restrictions imposed by entities controlling the harvesting of data (and in some circumstances also the entity controlling data repositories). In other instances AI as a Service ("AIaaS") providers or collaboration platforms include certain rights to use the data uploaded to their proprietary services and platforms.

Another observation is that a number of agreements tend to treat data as a uniform asset and apply a single, simplistic approach to all data exchanged between the parties in a data sharing arrangement. No strict legal definition of data exists, and without a precise contractual description, a number of issues are

left to interpretation, such as the potential for differing classes of data to be subject to differing levels of regulation. A generic uniform contractual approach of what is typically a significantly heterogenous landscape of data exchanged under a sharing arrangement may also lead to unforeseen effects, some with higher risk potential than others. Rather than applying a generic approach (i.e. a strict limitation on use and transfer to third parties), a trend we have seen of dividing the datasets into various categories according to their own specific regulation levels, seems to be a more practical way to deal with this issue.

Data trusts and trusted data intermediaries

In the previous version of *Responsible AI* we discussed the concept of the data trust. Since then there have been a number of initiatives undertaken to further refine the concept of the Data trust—that is to say an innovative legal structure providing for the independent stewardship of data. In particular the Open Data Institute in the UK published a report in 2019[9] explaining the lessons learned by it from three data trust pilots. As at the date of preparation of this update, the concept remains experimental. In the private sector, we are aware of a number of competing initiatives to develop a standard data interchange format which may involve a trusted data steward or intermediary acting as data guardian. For example, Engine B, an enterprise formed jointly by all of the "big 4" accountancy firms, with technological support from Microsoft and IBM has received £1.75m funding from Innovate UK to create a standardised data interchange format for accountancy and legal professional services.[10]

Alternative data sharing models

In this section we list some of the alternative data sharing models which are currently being proposed by various entities around the world.

For certain types of data it may be desirable to apply an open license regime based on the same principles found in the open source licenses used for open source software. These types of models are typically referred to as "data commons." Indeed, the Open Data Directive (2019/1024)[11] urges member states to encourage the use of open licenses in accordance with the Commission's guidelines and recommendations for standard licensing.[12] Some public datasets are made available based on the Creative Commons Attribution license and various national licenses are already in place.[13] Certain datasets will also be subject to mandatory license terms, such as the inbound license terms set out in article 14 of the Open Data Directive (2019/1024) which applies to high-value datasets (as defined in Annex 1).

In this regard, the Linux Foundation communities have developed data license agreements that enable sharing of data similar to that achieved with open source software. The result is a large scale collaboration on two licenses for sharing data under a legal framework called the Community Data License Agreement (CDLA).[14]

The CDLA Sharing license is based on the principles of "copyleft," ensuring that users of data governed by the CDLA Sharing license must ensure that downstream recipients enjoy the same conditions for using the data, including the right to modify the data and the obligation to share their changes to the data. The CDLA-Permissive license is similar to permissive open source licenses in that the publisher of data

allows anyone to use, modify and do what they want with the data with no obligations to share any of their changes or modifications.

Finally, the Open Knowledge foundation[15] has formulated a definition[16] of "open data" and published a license targeted at regulating open data related to culture, science, finance, statistic, weather and environment. Recipients of data received under such open licenses obtain irrevocable permission to redistribute, modify, separate and compile open data for any purpose. Such licenses also require that any further use of data is subject to similar and non-discriminatory terms. Certain conditions may be attached to the redistribution of open data by a party who has obtained data under an Open licence.

Principle 6
Open Data and Fair Competition

Organisations that develop, make available or use AI systems and any national laws that regulate such use shall, without prejudice to normal rules of intellectual property and privacy:

(a) foster open access to, and the portability of, datasets (where privately held), especially where such datasets are deemed significant and important or advance the "state of the art" in the development of AI systems;

(b) ensure that data held by public sector bodies are, in so far as is reasonably practicable, portable, accessible and open; and

(c) encourage open source frameworks and software for AI systems which could similarly be regarded as significant and important and advance the "state of the art."

AI systems must be developed and made available on a "compliance by design" basis in relation to competition/antitrust law.

1 Supporting effective competition in relation to AI systems

1.1 Governments should support and participate in international co-ordination (through bodies such as the OECD and the International Competition Network) to develop best practices and rigorous analysis in understanding the competitive impact of dataset control and AI systems on economic markets.

1.2 Governments should undertake regular reviews to ensure that competition law frameworks and the enforcement tools available to the relevant enforcement authorities are sufficient and effective to ensure sufficient access to necessary inputs, and adequate choice, vibrant rivalry, creative innovation and high quality of output in the development and deployment of AI systems, to the ultimate benefit of consumers.

2 Open data

2.1 Governments should foster and facilitate national infrastructures necessary to promote the portability of and open access to, datasets, especially those that are significant and important, to all elements of society having a vested interest in access to such datasets for research and/or non-commercial use to further advance the "state of the art" in relation to such technology and to ensure the efficacy of existing AI systems. In this regard, governments should give serious consideration to two-tier access models which would allow for free access for academic and research purposes, and paid-for access for commercialised purposes.

2.2 Governments should support open data initiatives in the public or private sector with guidance and research to share wide understanding of the advantages to be gained from open access data, the structures through which

datasets can be shared and exchanged, and the processes by which data can be made portable and suitable for open access (including API standardisation, pseudonymisation, aggregation or other curation, where necessary).

2.3 Governments should ensure that the data held by public sector bodies are accessible and open, where possible and where this does not conflict with a public sector mandate to recover taxpayer investment in the collection and curation of such data. Private sector bodies such as industry organisations and trade associations should similarly support and promote open data within their industry sector, making their own datasets open, where possible. The degree of relative influence that private sector organisations have on applicable markets should be assessed on a continuous basis by regulators.

2.4 Organisations that develop, make available or use datasets, especially those which could be regarded as significant or important or which could be regarded as advancing the "state of the art" are similarly encouraged to open up access to, and/or license, such datasets, where possible via chaperoned mechanisms such as Data Trusts.

2.5 Any sharing or licensing of data should be to an extent which is reasonable in the circumstances and must be in compliance with legal, regulatory, contractual and any other obligations or requirements in relation to the data concerned (including privacy, security, freedom of information and other confidentiality considerations). In addition, all stakeholders involved in such sharing or licensing should be very clearly identified in terms of legal roles, duties and responsibilities.

3 Open source AI systems

3.1 Organisations that develop AI systems are normally entitled to commercialise such systems as they wish. However, governments should at a minimum advocate accessibility through open source or other similar licensing arrangements to those innovative AI systems which may be of particular societal benefit or advance the "state of the art" in the field via, for example, targeted incentive schemes.

3.2 Organisations that elect not to release their AI systems as open source software are encouraged nevertheless to license the System on a commercial basis.

3.3 To the extent that an AI system can be subdivided into various constituent parts with general utility and application in other AI use-cases, organisations that elect not to license the AI system as a whole (whether on an open source or commercial basis) are encouraged to license as many of such re-usable components as is possible.

4 Compliance by design with competition/antitrust laws

4.1 Organisations that develop, deploy or use AI systems should design, develop and deploy AI systems in a "compliance by design" manner which ensures consistency with the overarching ethos of subsisting competition/antitrust regimes to promote free and vibrant competition amongst corporate enterprises to the ultimate benefit of consumers.

Endnotes

1 European Commission, "White Paper on Artificial Intelligence" (19 February 2020), COM (2020) 65 final, online: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

2 European Commission, "European Strategy on Data" (19 February 2020), COM (2020) 66 final, online: https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf.

3 European Commission, "European Commission Digital Strategy, A digitally transformed, user-focused and data-driven Commission" (21 November 2018), C(2018) 7118 final, online: https://ec.europa.eu/info/sites/info/files/strategy/decision-making_process/documents/ec_digitalstrategy_en.pdf.

4 EU, https://ec.europa.eu/digital-single-market/news-redirect/666643.

5 EU, Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information, online: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1561563110433&uri=CELEX:32019L1024.

6 See https://www.gov.uk/government/publications/smart-data-review/smart-data-review-terms-of-reference.

7 See https://blogs.microsoft.com/on-the-issues/2020/04/21/open-data-campaign-divide/.

8 These five principles are "Open" (making data relevant to important social problems as open as possible); "Usable" (investing in new technologies to make data more usable); "Empowering" (helping organisations generate value from their data); "Secure" (employing security controls to ensure operational security); and "Private" (helping to ensure privacy in data sharing collaborations).

9 https://docs.google.com/document/d/118RqyUAWP3WIyyCO4iLUT3oOobnYJGibEhspr2v87jg/edit.

10 https://www.businessleader.co.uk/engine-b-awarded-1-75m-innovate-uk-industrial-research-award/83356/.

11 Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information.

12 Recital 44, 64 and 66 of the Open Data Directive.

13 See for example the Open Government License for public sector information in the UK, online: https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/.

14 See Linux Foundation projects, https://cdla.io/.

15 https://okfn.org/opendata/.

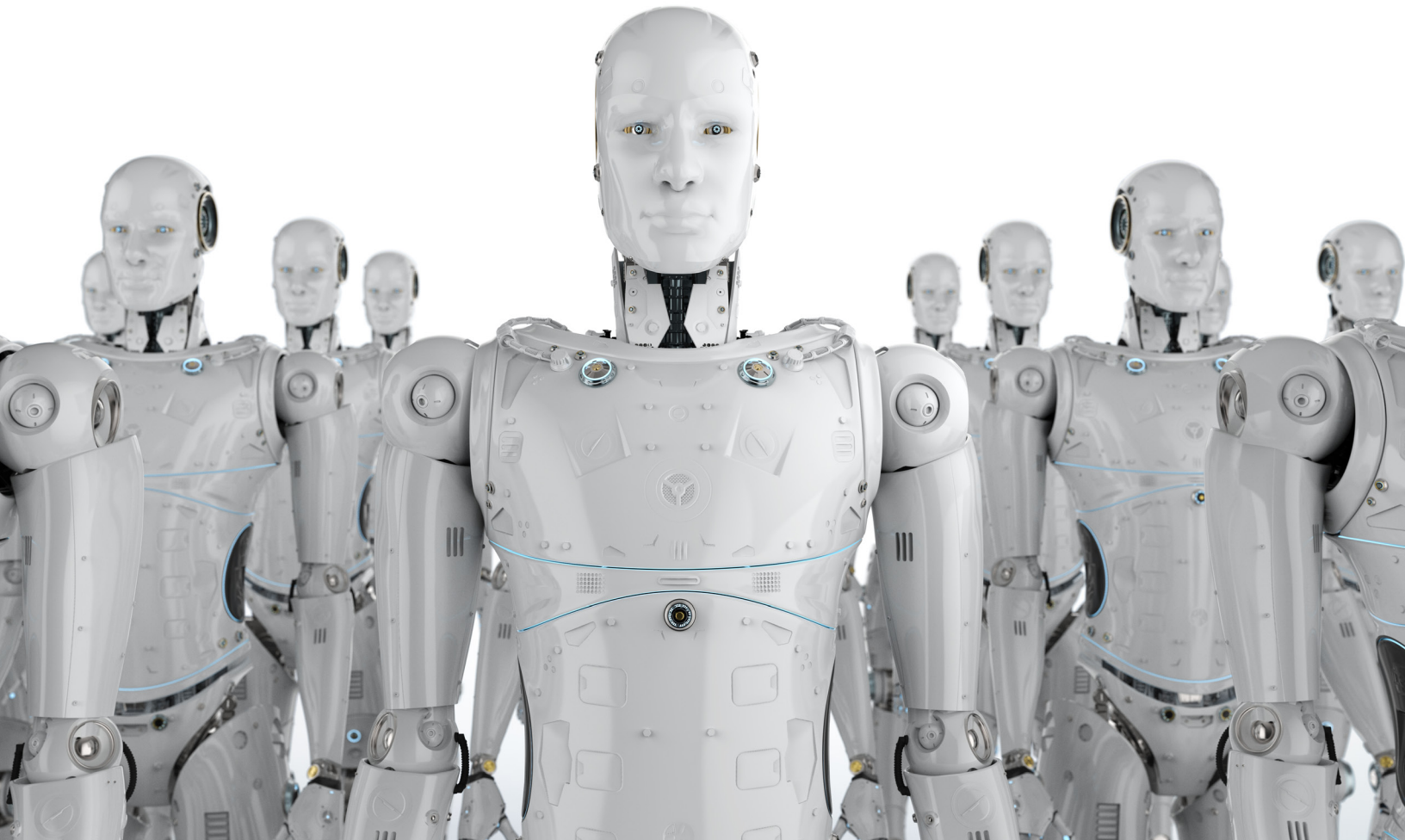
16 https://opendefinition.org/od/2.1/en/.

Responsible AI

A GLOBAL POLICY FRAMEWORK

Principle 7

PRIVACY



Principle 7 Commentary

PRIVACY

CHAPTER LEAD
Michael Peeters | DAC Beachcroft LLP, UK

Nicole Beranek Zanon | de la cruz beranek Attorneys-at-law Ltd., Switzerland

Dr Sonja Dürager | bpv Hügel Rechtsanwälte, Austria

Thomas de Weerd | Houthoff, Netherlands

Alexander Tribess | Weitnauer Partnerschaft mbB, Germany

Padraig Walsh | Tanner De Witt, Hong Kong

Belen Arribas | Belen Arribas Sanchez, Abogada, Spain

John Tomaszewski | Seyfarth Shaw, United States

Introduction

Since publication a year ago, we have seen a real groundswell of engagement in this important area by those responsible for regulating personal privacy. This has led to a wealth of papers on AI issued by privacy regulators and related bodies from around the world. Our update review of the Principles (see below) has been informed in large part by these various papers and the themes that emerge from them, as well as jurisdictions which have updated their privacy laws in the last year or so. In most cases the changes made have been fine-tuning of wording as thinking has matured.

It should also be borne in mind that this proliferation of interest by privacy regulators to review the implications of AI has led to a disproportionate emphasis in this area: it is sometime easy to forget that not all AI makes use of personally identifiable information—and compliance should extend beyond privacy. There is a real risk, particularly with powerful data regulators in Europe, that they "fill the AI compliance vacuum" to the exclusion of other factors. That said, this update will now focus on privacy and AI!

Key Developments in Privacy since Original Publication

The EU White Paper on Artificial Intelligence (and subsequent moves towards regulation)

As mentioned in earlier Chapter updates, in February 2020, the European Commission published its White Paper on Artificial Intelligence which comments on privacy issues in AI.

The White Paper sets out various scenarios in which AI could adversely affect the right to privacy, emphasizing the risk that AI would increase the possibilities to track and analyse the daily habits of people, e.g. by state authorities and other entities for mass surveillance and by employers to observe how their employees behave. The Commission considers that, by analysing large amounts of data and identifying links among them, AI may also be used to retrace and de-anonymise data about persons, creating new personal data protection risks even in respect to datasets that, *per se,* do not include personal data. AI may also be used by online intermediaries to prioritise information for their users and to perform content moderation. These developments are, in the opinion expressed by the Commission, likely to result in a breach of EU data protection and other rules (such as, for instance, the GDPR). The Commission states: *"The processed data, the way applications are designed and the scope for human intervention can affect the rights to free expression, personal data protection, privacy, and political freedoms"* (p. 12).

However, when it comes to the principles for future legislation, specifically tailored to the field of AI, the Commission holds that, when setting up a new regulatory framework for AI, the legislator should adhere to a risk-based approach. The objective of any regulation concerning AI application should be to balance out the risks for the rights and freedoms of natural persons (as outlined above) whilst not being excessively prescriptive so that it could create a disproportionate burden, especially for SMEs. The Commission endorses, for high-risk AI applications, a set of requirements addressing the following key features, subject to further specification to ensure legal certainty: *"training data; data and record-keeping; information to be provided; robustness and accuracy; human oversight; specific requirements for certain particular AI applications, such as those used for purposes of remote biometric identification"* (p. 18 et seq.).

On the basis of this assessment in the White Paper, in Spring 2020, the European Commission initiated a public consultation concerning a first initiative for a future European legal framework in the field of AI; the aim being to ensure that AI is safe, lawful and in line with EU fundamental rights (to thereby stimulate the uptake of trustworthy AI in the EU economy). The consultation closed in June and the Commission, according to its communications, is still considering different approaches, stretching from soft-law merely promoting industry initiatives for AI to EU legislative instruments establishing mandatory requirements for all or certain types of AI applications. The current time-line prospects an adoption of a Commission decision for early 2021.

The risk-based approach and the Commission's statements as to protection of the economic interests of SMEs sound familiar. The Commission has used similar language as regards the GDPR. However, in practice, the GDPR has been widely considered by companies to impose excessive obligations and documentation duties on SMEs also for their day-to-day business. The constitutional rights of EU citizens to privacy and data protection will pave the way for future legislation also in the field of AI.

In our view, this GDPR approach should be reconsidered and a solution be sought that includes a more targeted and precise framework for companies using AI-driven applications. This could be achieved, for instance, by combining such general clauses with bans: the regulatory framework could contain certain blacklisted AI processing activities, and hence, a list of practices that tend to threaten or violate the data subject´s rights (e.g. certain profiling measures).

Europe wide privacy regulators—an ongoing debate

As mentioned above, a number of privacy regulators across Europe have issued guidance in this area in the last year or so: the extent of this illustrates how there is a disproportionate emphasis on privacy regulation in the AI space to the expense of other ethical/legal issues. That said, we set out just a couple of examples to illustrate the trends here:

- **The proliferation of non-enforcable guidelines by privacy regulators:** for example, the Spanish Data Protection Authority (AEPD) published in March 2020 guidelines which review the most important matters that must be taken into account when designing products and services that carry out data processing using AI. The guidelines stress that AI generates many doubts in relation to regulatory compliance, in particular, regarding the rights of the data subjects. The document concludes that quality and privacy guarantees need to be applied. This illustrates the debate across European privacy regulators like with other regulators, the Spanish regulator is tasked with, inter alia, establishing guidelines for stakeholders (in this case developers and vendors of AI solutions).
- **The use of privacy regulation to tackle the use of AI in Automated Decision Making:** for example,
  - in the **UK**, the Information Commissioner's Office (ICO) has published draft guidance on Explaining AI Decisions targeted at technical, development and senior management teams.
  - in some jurisdictions this has now reached specific legislation building upon the existing GDPR high level provision: the new **Swiss Federal Act on Data Protection**[1] will introduce (in early 2022) various specific obligations with regard to AI-based, automated individual decisions. According to Art. 21, *"a person responsible shall inform the data subject of a decision based exclusively on automated processing which entails a legal consequence for him or her or significantly affects him or her (automated individual decision). On request, he shall give the data subject the opportunity to state his position. The data subject may request that the decision be reviewed by a natural person."*

Whether a decision (which has to be of a certain complexity) is based solely on automated processing is the subject of detailed definition but effectively depends on the absence of any natural person in the assessment of the content and making of a decision based on this assessment. The data subject does need to be informed only if the decision entails direct legal consequence for the data subject or significantly affects him or her; and the responsible person must also inform the data subject about profiling if this leads to such a decision. The controller must give the data subject the opportunity to make his point of view known if he so requests. In particular, he shall be given the opportunity to express his views on the outcome of the decision and, where appropriate, to ask how the decision was reached. This does not apply, "if the decision is directly related to the conclusion or performance of a contract between the responsible person and the data subject and his request is granted or if the data subject has expressly consented to the decision being automated."

USA: The federal approach to AI and California's new privacy law

The United States has historically maintained a fairly "hands off" attitude to technology innovation from a regulatory perspective. This is at both the Federal and State levels. As a consequence, there is a "patchwork" of sector-specific laws and regulations which address the concepts of privacy and personal

autonomy. However, there have been some recent developments which see the US (or at least parts of it) moving in the direction of the rest of the world with respect to data protection—most significantly in California with the CCPA—leading to a question as to whether these new regulatory developments restrict AI deployment and how?

While not directly applicable to the "use" of AI, these regulations may well have a direct impact on the "development" of AI. This is because machine learning engines need a tremendous amount of data to be able to tune their algorithms; and this automated use of data for purposes outside the original purpose of collection (i.e. Secondary Use) is going to be a barrier to the development of AI as, while Secondary uses have historically been permissible in the US, the approach is shifting away from such permissible secondary uses, toward a more narrowly focused primary use approach like exists in the EU. As such, AI development may become an unwitting victim of the evolving privacy regime in the US.

California

The California Consumer Privacy Act (CCPA) was enacted in 2018 in response to a grassroots initiative to bring GDPR-style law to the US. While the GDPR was one of the motivations for the actions of California, it is important to note that there are some very distinct and different approaches to data protection between the GDPR and the CCPA. However, both approaches will have a material impact on AI and its development over time.

The CCPA does include a new set of "privacy rights" (most US law previous to the CCPA dealt with security and not really privacy). Similarly to the GDPR, the CCPA now includes the rights of notice, access, deletion, and objection of processing. However, this is generally where the similarities stop:

- the CCPA permits a broader set of uses for data—including use of data for AI development;
- when such development is not directly related to the primary purpose for collection so long as such secondary purpose is disclosed in a notice at collection;
- under the CCPA, AI-based processing is not *per se* prohibited.[2]

In short, the CCPA places very few GDPR-style restrictions on data processing which could affect AI development or use. However, California is not the only organ of the US which is looking at AI use and it attendant risks.

Federal Trade Commission

Regulation affecting AI can come from sources which originally never contemplated regulating this kind of activity. The US Federal Trade Commission (**FTC**) has historically used its powers under the twin doctrines of "unfairness" and "deceptiveness" enshrined in Section 5 of the FTC Act to regulate practices which are not generally seen as purely "commercial." It is these twin doctrines which the FTC would use to try to regulate AI at both its developmental and deployment stages. For example, if AI development didn't take into account bias inherent is certain data-sets, the processing could be viewed as unfair. If the AI

processed data from third parties where there was no disclosure that 1) it was an AI doing the processing and 2) what the sources of such data are, that could be considered deceptive.

The FTC also has a specific law in its arsenal to regulate "big data" and AI. The US Fair Credit Reporting Act (**FCRA**) is being used by the FTC in much of the same way that Section 5 of the FTC Act has been used: applying a law which is not originally designed to regulate technology to technology.[3] The FCRA has a very broad definition of what constitutes a "Consumer Reporting Agency"—which may well apply to AI uses. In short, if data processing is done for a "permissible purpose" under the FCRA (e.g. determining eligibility for employment, evaluating suitability for a commercial transaction, etc.) then such data processing will likely fall under the FCRA's system of rights and obligations (which look more like the GDPR than the CCPA does).

Additionally, the FTC has brought cases alleging violations of the laws involving AI and automated decision-making, and have investigated numerous companies in this space. The FTC's scrutiny into "big data" (and thus AI) has been apparent since at least 2016 when they published their report titled "Big Data: A Tool for Inclusion or Exclusion?"[4]

As a consequence, much of the US regulatory activity around AI is not at the state level, but is being spearheaded by the FTC who is using the FCRA as its "big data" law. As FTC Act Section 5 morphed from "deceptive trade practices" into "privacy," the FCRA is morphing from "credit report protection" to "AI regulation." Further, since the FCRA is a federal law, it has broad applicability across all 50 states—which is an inherent limitation of the CCPA.

Asia

Asia has seen several jurisdictions review their privacy law and make proposals for legislative reform. This is the ripple effect of GDPR. Many businesses have effectively adopted GDPR standards in their business as a consequence of international data flows and the corresponding legal consequences. This has led to a perception of a higher standard for the international norm, and a lower resistance to introducing change in domestic laws. We have seen new laws enacted or introduced to legislative chambers in Thailand, India, South Korea and Japan, and proposals for new laws announced in Hong Kong, Singapore, and Malaysia. China also introduced privacy related provisions in its Civil Code, and has published a draft data security law. Asia can no longer be seen as lagging on personal data and privacy law. As part of this evolution in privacy law across major Asian nations, we have seen an initial development in certain areas of relevance to AI: what follows is a 'round-up' of these.

China—the landmark first 'Civil Code' is expected to be effective from January 2021. It covers a wide spectrum of rights including many new laws relating to personal data, some of relevance to AI:.

- For the first time, there are exemptions from liability available to lawfully handle personal information including personal data from public sources. AI users should exercise caution when relying on automatic data collection to process publicly available information. If the information was not lawfully

uploaded (e.g. photos taken without the authority of the person), then use of that personal information still constitutes an infringement of the person's rights.

- Processing must be lawful, justified, necessary and not excessive and personal data must not be disclosed or amended without consent, unless it cannot identify any individual.
- Certain individuals' rights were also introduced; for example, the right to obtain access to personal data kept by an organisation, to correct personal data or to request deletion of the data.

AI users in PRC should therefore exercise care when collecting and handling personal data in China. Appropriate measures should be in place to ensure only lawful personal data are collected, and good internal data security measures should be implemented.

Hong Kong—the Privacy Commissioner published a report in 2018 to advocate an ethical accountability framework for data stewardship—not confined to personal data but rather focus on any data and data-driven activities. The report is non-binding; however, it is supported by leading businesses and business associations in Hong Kong, and therefore it may not be surprising to see new regulations similar to this report being introduced. The report introduced the framework of enhanced accountability for data-stewards (in contrast to data-custodians), and provided validation methods such as the new ethical data impact assessment (EDIA) and the Process Oversight Model—both of potential relevance to an AI deployment. There was no requirement to appoint an AI Ethics Officer and compliance should be done through internal or external audits.

Singapore—the Singapore government published the second edition of its Model Artificial Intelligence Governance Framework. It advocates establishing a central coordinating body with relevant expertise, and proper representation from across the organisation (known as the "human-over-the-loop") to encourage organisations to have sufficient control over technology. However, it stops short of advocating an AI Ethics Officer. The framework introduced the implementation and self-assessment guide for organisations (ISAGO) and the compendium of use cases. The ISAGO is designed to be a guide for organisations when implementing the framework and to work with their existing AI governance; and the compendium is a set of case studies for references. Three new measures were also introduced to help organisations to enhance the transparency of the algorithms used in AI models. The framework does not directly address personal data issues but refers to the requirement that organisations must understand the lineage of data they are processing. It is not intended to limit the scope to a narrow personal data focus, but instead to a broader ethical data framework. The government does not intend for the framework to be binding but is encouraged to be adopted by organisations, as it will assist in compliance with the existing Personal Data Protection Act 2012. AI users shall monitor the development of the model framework closely as it is expected to be refined by the government following feedback from the industry. In recently announced proposals for legislative reform, there may soon be an exception to the requirement to obtain data user consent if the personal data is collected, used or disclosed where the legitimate interests of the data controller and the benefit to the public is greater than the adverse effect to the data subject. This may provide some safe harbour for AI users.

Japan—regulators enacted amendments to their Act on the Protection of Personal Information (APPI) in June 2020. The changes are expected to come into effect in spring 2022 and will bring Japanese personal

data regulation closer to the GDPR. This includes introduction of new concepts such as "personally identifiable information" (which can potentially cover cookies) and "pseudonymised information," as well as new rules on data processing storage. It is anticipated that the amendments can promote the usage of data in the context of AI by providing appropriate regulatory directions. In addition to the changes to the APPI, the Ministry of Economy, Trade and Industry published a white paper in July 2020 that addressed the need for new governance models with respect to big data, IoT, AI and other digital technologies. Therefore, there may be more changes to privacy regulation in the near future in Japan that may impact AI users.

South Korea—several privacy laws are being amended effective August 2020. This includes the Personal Information Protection Act (PIPA) and the Credit Information Protection Act (CIPA). Similar to the GDPR, the concept of "pseudonymised information" is being introduced. Other changes include clarification and relaxation on the constraints on permitted usage and transfer of customer data. The Personal Information Protection Commission (PIPC) will be vested with the power to monitor, make policy and regulate the practice. Although many ambiguities arising from the amendments remain to be clarified, it is clear that there is a trend to allow more big data-based services to take place. AI using organisations should therefore monitor the development of the legislation closely to ensure their businesses are in compliance.

India—Unlike other countries in South East Asia, India does not have any comprehensive data privacy law: however, this is expected to change with the introduction of the Personal Data Protection Bill 2019 (PDP Bill). The PDP Bill is broadly based on the GDPR and has some provisions of relevance to AI including:

- a broad requirement to delete personal data after processing;
- the possibility of a sandbox to encourage innovation businesses in AI, machine learning or other emerging technology.

Rise of biometrics and facial recognition

The growing use of AI in biometrics and facial recognition is well documented (and referred to in other Chapters) but is much more of a privacy issue than when we originally published, especially since the advent of the Covid-19 pandemic (with tracing systems). There is no doubt some applications of Automated Facial Recognition (AFR) are beneficial to society in some cases (mainly diagnosis of health issues) but there are concerns: issues include; how consent may be practically obtained; whether a pseudonymised (such as those for customer profiling in retail applications) system is subject to the same consent requirements; whether the systems are accurate; and where state surveillance may be allowed.

Cases are now beginning to examine the privacy issues behind the use of this technology. A recent example is the UK South Wales Police Case[5] where the UK's High Court considered the lawfulness of the use of AFR by the South Wales Police during its policing operations. As part of a trial, AFR was used with surveillance cameras in public spaces (e.g. at football and rugby matches and music festivals) which captured images of members of the public, that were then compared to digital images of individuals on a 'watch-list'. In the event of there being no match between the surveillance image of an individual and an image within the watch-list, all biometric data about that individual was immediately deleted, with the CCTV footage being retained as per the usual retention period. The Claimant's image had been captured on

two occasions and did not appear on the 'watch-list' (so his biometric data was deleted). The claim was brought mainly under the UK data privacy law[6] but the Court rejected the claim: a significant factor in leading the Court to conclude that the use of AFR was lawful was the fact that there was always a review by a police officer casting a "human eye" over the AFR software's identification decisions.

Another interesting recent case is the Swedish School Board decision[7] (August, 2019) where a High School in Sweden was found to have violated the GDPR[8] by the Swedish Data Protection Authority when it co-developed and tested a facial recognition system to track student attendance. Despite getting the consent of students and their guardians, it was felt that due to the power imbalance between the students and school staff as well as the sensitive and excessive nature of data being collected, there was insufficient legal basis for data processing.

Overall this is an area to monitor going forwards; clearly there is much debate ongoing as to the use of AFR in state surveillance and the human rights issues. Equally work needs to be done to improve private sector use (for example in the retail sector).

Revisions to Principle 7

Changes to Principle 7.1

The changes to Principle 7.1 highlight some of the most important factors in the challenge we face of finding a balance between the protection of personal data and the opportunities that come with the increased use of AI systems.

Key to this is the concept of "privacy by design" which calls for the implementation of appropriate technical and organisational measures when developing or using AI systems; the underlying idea being that privacy principles need to be taken into account at every step during the development and use of AI systems in order to prevent breach of data protection law. These measures can include minimising the processing of personal data, pseudonymising personal data as soon as possible, or transparency with regard to the functions and processing of personal data (as per Recital 78 to the GDPR). They may also include avoiding the processing of personal data completely by making use of anonymous data (potentially synthetic data) for AI training purposes—an approach recently recommended by German DPAs.[9]

There should be always a balance between the benefit and threats of AI. This will vary depending on the application in question as, for example, recommendations for music titles are far less critical than identification of potential criminal suspects, with a resulting different level of acceptance of AI deployment. This distinction has now begun to flow into DP laws; e.g. in the new Swiss privacy law the data controller must inform the data subject of any decision taken solely on the basis of automated processing of personal data (including profiling) but only where this has legal implications for the data subject or significantly affects him/her (Art. 19 of the draft Data Protection Act which is equivalent to Art. 22 GDPR).

Changes to Principle 7.2

This Principle refers to the responsibilities of organisations developing, deploying and using AI systems to implement operational safeguards to protect privacy. However, recent months have shown that there is uncertainty as regards the allocation of responsibilities between different economic operators in the supply chain of developing and using AI systems (see also EU Commission White Paper on AI—above).

In the context of AI systems, the fact is that no company can create an AI solution alone, and vendors increasingly must form strategic partnerships that give them access to all the necessary complementary technologies and data. In particular, if the developer of the algorithms and the entity responsible for continuous machine learning that use the input data, are different from the entity that sells the AI based end product that creates the output data (which may then be used again as input data), then this is a complex factual situation for the allocation of responsibilities. Concepts of who controls machine learning (alone or jointly with developers of AI algorithms), who deploys AI algorithm in the end user products and who collects the data from the consumers, play an essential role in the application of the data protection rules, since they determine who shall be responsible for compliance with privacy rules, and particularly, how and towards whom data subjects can best exercise their rights in practice.

Principle 7
Privacy

Organisations that develop, make available or use AI systems and any national laws that regulate such use shall endeavour to ensure that AI systems are compliant with privacy norms and regulations, taking into account the unique characteristics of AI systems, and the evolution of standards on privacy.

1 Finding a balance

1.1 There is an inherent and developing conflict between the increasing use of AI systems to process private data, especially personal data; and the increasing regulatory protection afforded internationally to personal and other private data This protection typically applies principles of purpose limitation, data minimisation and storage limitation.

1.2 Governments that regulate the privacy implications of AI systems should do so in a manner that acknowledges the specific characteristics of AI and that does not unduly stifle AI innovation.

1.3 However, governments should foster the privacy principles, in particular of purpose limitation, for personal data within the use of AI systems.

1.3 Organisations that develop, make available and use AI systems should analyse and constantly check their current processes to identify whether they need be updated or amended in any way to ensure that the respect for privacy is given as a central consideration. This includes consideration as to whether and to want extent AI systems actually require the processing of personal (as opposed to, e.g. anonymous) data.

2 The operational challenges ahead for AI users

2.1 AI systems create challenges specifically in relation to the practicalities of meeting of requirements under a number of national legislative regimes, such as in relation to consent and anonymisation of data. Likewise AI systems create challenges as to data subject rights and legal certainty for all parties involved. Accordingly, organisations that develop, deploy or use AI systems and any national laws that regulate such use, shall make provision for alternative lawful bases for the collection and processing of personal data by AI systems, such as a rightful use of the input and output data.

2.2 Organisations that develop, deploy or use AI systems should identify the level of responsibility when they use input or output data for AI systems (e.g. to avoid unlawful discrimination). Organisations should then consider the resulting consequences and obligations, including implementing operational safeguards to protect privacy such as privacy by design principles that are specifically tailored to the specific features of deployed AI systems.

2.3 Organisations that develop, deploy and use AI systems should appoint an AI Ethics Officer, in a role similar to Data Protection Officers under the GDPR, but with specific remit to consider the ethics and regulatory compliance of their use of AI.

3 AI as a tool to support privacy

3.1 Although there are challenges from a privacy perspective from the use of AI, in turn the advent of AI technologies could also be used to help organisations comply with privacy obligations.

Endnotes

1 https://www.parlament.ch/centers/eparl/curia/2017/20170059/Schluzssabstimmungstext%203%20NS%20D.pdf.

2 It should be noted that the California Privacy Rights Act—a successor to the CCPA—is contemplating the inclusion of AI-only processing in its prohibited activities.

3 The FTC's guidance on AI and automated decision making follows the basic OECD Privacy Principles. The one interesting aspect of the guidance is the aspect of scientifically sound principles being necessary in the modeling used for automated decision making. This is a specific application of the OECD principles. See https://www.ftc.gov/news-events/blogs/business-blog/2020/04/using-artificial-intelligence-algorithms?utm_source=govdelivery.

4 https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf.

5 R (Bridges) v Chief Constable of South Wales Police and Others [2019] EWHC 2341.

6 Data Protection Act, 2018.

7 https://www.imy.se/globalassets/dokument/beslut/facial-recognition-used-to-monitor-the-attendance-of-students.pdf.

8 Article 5 (1) c), Article 9, Article 35 and Article 36 GDPR.

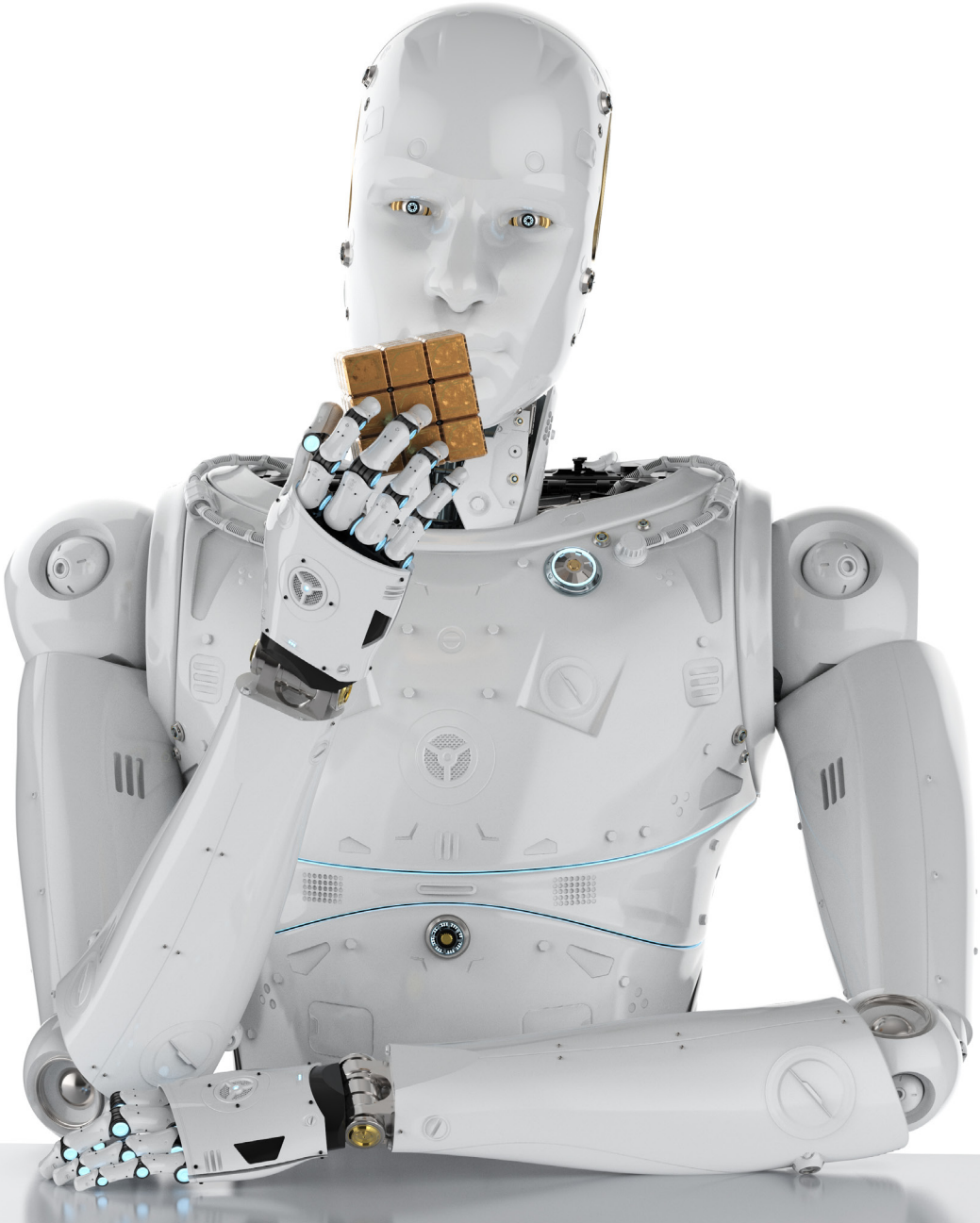
9 Resolution of the 97th Conference of the Independent Federal and State Data Protection Supervisory Authorities of Germany April 3, 2019—Hambach Declaration on Artificial Intelligence, a whitepaper on AI issued by the German DPAs.

Responsible AI

A GLOBAL POLICY FRAMEWORK

Principle 8

AI AND INTELLECTUAL PROPERTY



Principle 8 Commentary

AI AND INTELLECTUAL PROPERTY

CHAPTER LEAD
Susan Barty | CMS, UK

Segolene Delmas | Lawways, France
Marco Galli | Gattai, Minoli, Agostinelli, Partners, Italy
Licia Garotti | Gattai, Minoli, Agostinelli, Partners, Italy
Julian Potter | WP Thompson, UK
Rory Radding | Mauriel Kapouytian Woods LLP, USA
Gilles Rouvier | Lawways, France
Alesch Staehelin | TIMES Attorneys, Switzerland

Introduction

Since the publication of *Responsible AI: A Global Policy Framework,* there have been a number of recent developments relating to the interface between AI and IP. One such development is the attempt to apply for patent protection in the US, the EU and the UK, based on inventions where the inventor was designated as DABUS—an AI machine. The US Patent and Trade Mark Office (USPTO), the European Patent Office (EPO) and the UK Intellectual Property Office (UKIPO) decisions in the DABUS case, all reached the same conclusion that an AI machine was not capable of being an inventor under respective patent legislation and rules. The question of the patentability of the subject-matter alleged to be invented by DABUS was not considered as the application failed on formal grounds before entering substantive examination.

Also, the World Intellectual Property Organization (WIPO) has been seeking to develop, through an open process, a list of issues concerning the impact of AI on IP, to form the basis of future structured discussions and, it is hoped, some consensus on how to develop a policy for IP and AI. WIPO held its first Conversation on IP and AI in September 2019 to support this process, inviting member states and other interested parties to provide comments and suggestions. WIPO then launched a public consultation process on artificial intelligence (AI) and intellectual property (IP) policy (the WIPO Consultation), inviting feedback on a number of issues which had been identified as the most-pressing questions likely to face IP policy makers as AI increases in importance. The WIPO Consultation was launched on 13 December 2019 with a Draft Issues Paper[1] and over 250 responses were received.[2]

The issues identified by WIPO for discussion were originally divided into six basic areas: Patents, Copyright, Data, Designs, Technology Gap and Capacity Building, and Accountability for IP Administrative Decisions.

WIPO has, on 29 May 2020, published a revised issues paper ("the WIPO Revised Issues Paper") (the paper itself is dated 21 May 2020) on IP and AI,[3] taking into account the comments received as part of the WIPO Consultation, as part of its ongoing conversation with stakeholders on the intersection of AI and IP policy. Another WIPO Conversation on IP and AI took place from 7-9 July 2020.[4]

The WIPO Revised Issues Paper[5] develops on the previously identified six basic areas. It picks up on the need for agreed definitions when talking about "AI," "AI-generated," "autonomously generated by AI," "AI assisted" etc, the need to address areas of trade mark law that may be impacted by AI and the impact of trade secrets in terms of providing protection. The revised list of issues is:

- Glossary (Definitions)
- Patents (Inventorship and Ownership; Patentable Subject Matter and Patentability Guidelines; Inventive Step or Non-Obviousness; Disclosure; General Policy Considerations for the Patent System)
- Copyright (Authorship and Ownership; Infringement and Exceptions; Deep Fakes; General Policy Issues)
- Data (Further Rights in Relation to Data)
- Designs (Authorship and Ownership)
- Trademarks
- Trade secrets
- Technology Gap and Capacity Building
- Accountability for IP Administrative Decisions

This chapter update will only review certain of the issues identified by WIPO, where there was less focus in the first edition of *Responsible AI* and/or recent developments, namely deep fakes, the possible development of new rights in relation to data and issues relating to trade marks, and trade secrets. This update also reviews important decisions in different jurisdictions as to whether an AI system can be considered an "inventor" for the purposes of a patent application.

Patents—the DABUS cases

What if a machine (sometimes referred to as an AI or Machine Learning system) outputs a result that may be considered a new process or product—should the patent office recognise the "the machine" (or the AI system) as the inventor? This is the key question of the "DABUS" case that has made its way through patent offices in the US, Europe and the UK. The case poses several fundamental questions that might change well-established patent law principles.

Back in 2018, a group of AI aficionados filed two patent applications and designated an AI system as the inventor of each. However, in some jurisdictions where these applications were filed, the respective patent laws only recognise individuals or natural persons as the inventor. The patent applications were filed respectively for inventions for a specially shaped container lid designed for robotic gripping and a flashlight system for attracting human attention in emergencies. They were alleged to have been "created" by an AI system called DABUS (an acronym for "Device for the Autonomous Bootstrapping of Unified

Sentience") that was built by Mr. Stephen Thaler, founder and CEO of Imagination Engines Inc. of St. Charles, Montana, US.

DABUS is the result of over a decade of development; it was built to absorb data about a range of topics, including flashing light patterns and fractal geometry and, the key issue, to conceive ideas for products it had not seen before. The applications claimed that Stephen Thaler had no background whatsoever in developing container lids or flashlight systems, did not conceive of those two products and did not direct the machine to invent them. Accordingly, it was argued that it would be wrong to list Mr. Thaler as the inventor.

The argument seems to be akin to making the following analogy: If a natural person educates another individual in technology and the individual moves forward and independently develops something new, then the original educator does not become the inventor—when a machine is inventing instead of an individual it should be no different. Such an analogy does however require acceptance of the notion that a machine can invent but as the analogy is directed to the question of whether or not a machine can invent, the answer to the question should not be in the analogy. Additionally, a new notion appears to be being introduced, that of a machine being "educated." It may be better to phrase the question as being: if a machine has been configured to adapt its configuration depending upon data input to the machine, should a result produced by machine not directly linked to the data inputs be considered an independently created output of the machine? If the answer to the previous question is yes, should the machine be considered to be a creative entity?

So far, the applicants (who included Stephen Thaler) have filed patent applications with the UK Intellectual Property Office, the European Patent Office, the Israel Patent Office and the US Patent and Trademark Office (USPTO)—in each case, they have listed DABUS as the inventor.

In August 2019, the US patent office stated that the applications would not be considered unless the applicants listed the inventors involved by their legal names.[6]

In their response, the applicants requested that the US patent office should recognise DABUS as the inventor because there was no human inventor. The applicants also asked that Stephen Thaler be granted ownership IP rights to the inventions, as DABUS' employer or its successor in title. They claimed that the applications represented a test case that had implications for fairness, innovation and business certainty. It is unfair, so they argued, for people who do not themselves invent to be acknowledged in the same way as people who do. In addition, if companies see risks or impediments in seeking patents for AI generated inventions, they might be less inclined to use AI.

The USPTO issued its decision on 22 April 2020 stating that that AI systems cannot be listed or credited as inventors on a US patent; an "inventor" under current US patent law can only be a "natural person."[7]

In early December the UK Intellectual Property Office also refused the two UK DABUS patent applications on the basis that DABUS was not a "person" as required by the Patents Act and so could not be considered an inventor of a patent, despite arguments that the failure to acknowledge DABUS as "the actual devisor of the invention" would mislead the public.[8]

In January 2020, the European Patent Office (EPO) also rejected the two patent applications for DABUS.[9]

The applicants argued before the EPO that the impossibility for DABUS to provide its consent or to exercise its rights—such as moral rights—would not affect its designation as the inventor. Since the inventor is the one conceiving the invention, only DABUS meets this requirement. Designating an entity, or person, other than DABUS would have meant providing the EPO with untrue information, thereby undermining the public's right to know the actual inventor. Moreover, although not every natural person can validly provide his or her consent legally, this does not prevent him or her from being designated as the inventor.

Such arguments did not convince the EPO examiners, who refused the patent applications as they failed to meet the requirement of the European Patent Convention that the designated inventor must be a human being.

Since a patent office has no duty or power to decide whether AI systems shall be granted legal personality, the EPO emphasised that such a decision needs to be left to lawmakers or, at least, to case law.

In other words: If AI systems (such as IBM's Watson or Google's DeepMind) become not just competitive with a human inventor, but outperform the human inventor, companies would want to be using the AI systems in R&D. However, if patent offices refuse to grant patents to AI systems, then the companies might lose interest in developing AI systems further. As an example, a life science company using AI systems (or cognitive engines) to winnow-down chemical compounds that could be used in developing new drugs, could refrain from making enormous investments in these AI systems if the granting of patents remains more than questionable. Alternatively, the AI system could be considered to be merely a tool, the results of which are utilised by human beings to create innovations which may then be patentable subject matter.

Why is this case so important? Patents that list the wrong inventor or exclude an inventor can be challenged and deemed unenforceable. In a way, the DABUS case has some parallels with the "monkey selfie" case[10] where the copyright was not granted to the monkey and where there has been much debate as to whether the nature photographer who set up the camera could claim copyright ownership. Nevertheless, where there is no human connection in the development or creation of something new, individual and/or original, most contemporary copyright and patent regimes do not grant intellectual property rights to anyone.

The legal uncertainty surrounding the DABUS case could lead to a great deal of frustration in the developing industry in the field of patent law, where huge investments often precede the patenting process. Thus, many industry representatives are demanding that, regardless of whether an AI system or the person behind it gets credited as the inventor, the patent should be awarded, unlike in the monkey case (where the granting of copyright was at stake): The biggest industry concern is that no patent will be granted at all.

These decisions highlight the real need for wider debate as to the issue of AI systems whose output may be considered inventions. There is a real need for judicial and legislative branches of the law—across jurisdictions—to decide whether or not, and if yes, in what circumstances, a patent system should recognise AI systems as inventors.

Such a debate is also important because it should lead to an examination of how AI and machine learning systems work at a deep technical level which may inform the debate on the concept of invention and inventorship by machines. After all, the machine only does what it does because of its program and parameters. To that extent the machine is deterministic; start with the same initial state and input the same data in the same order and you will get the same result as before, unless there is some perturbation to the initial state or the machine is programmed to behave differently for subsequent iterations, possibly depending on the result of a previous iteration or iterations; although that may be considered to be a change in the initial state of each iteration.

Nevertheless, in August 2019, the US patent office sought comments from the public on the patenting of Artificial Intelligence Inventions.[11] These comments were made available for public inspection in March 2020[12]. The majority of the publicly available comments concluded that, at least for now, the emergence of AI technology does not justify modifying the limitations of named inventors to natural persons.

Patent issues are identified up front in the WIPO Consultation, where the first five issues identify issues which need to be considered in relation to patents and AI, highlighting the need to "encourage the investment of human and financial resources and the taking of risk in generating inventions that may contribute positively to the welfare of society." It is to be hoped that the outcome of the WIPO Consultation can bring some clarity to these issues.

If inventions made by an AI cannot be patented, this may lead organisations to seek trade secret protection for AI-made inventions. However, the prospect of keeping secret forever the best way to proceed in a field and industry, so the "progress of science and useful arts" (as per the US constitution) would not occur does not seem to be the most appropriate way of proceeding.

Patents—disclosure

One of the issues highlighted in the WIPO Revised Issues Paper is that of disclosure.[13] It is important to consider how disclosure would be handled in relation to AI and algorithms of machine learning which will be changing over time. The patent application must disclose the invention in a way that enables a person having ordinary skills in the art to implement the invention. Only sufficient disclosure grants to the inventor an exclusive right over it.

If the patented item is an AI-system, or a machine learning algorithm changing over time, disclosing the initial algorithm would not be likely to disclose the invention in a manner which is sufficient for the invention to be implemented by a person having ordinary skill in the art. Therefore, the issue of sufficient disclosure should be addressed providing, together with the initial algorithm, also a description of how the model is trained, including the relevant training data. Nevertheless, the applicant may not be willing to make these datasets available to its competitors, as they may "free ride" on them to train a different AI model. It may therefore happen that the description of how the model is trained is included in the patent application, whilst the training data are omitted.

Alternatively, it may become more likely that examples of the training data, rather than the complete data set, would be disclosed, one example of which is the UK patent application for Blippar.[14] The invention was an open set object recognition system and examples of the data set were provided to illustrate what was required to train the system. In the same way, a system which is configured to analyse input from a wide and varied data set and identify pattens and relationships between the data elements may produce a result that is new but nevertheless is a consequence of the system's programming and the data input to it. The complexity and wide range of data elements may give the appearance of cognition.

At the present date we are not aware of any existing case law that has refused to grant an AI-related patent application due to lack of training data.

Patents—examination issues

The consideration of the non-obviousness requirement to AI-generated inventions raises interesting issues as to whom the invention should not be obvious. The simple question is would a person skilled in the art find the "invention" obvious; but who should be considered as the person skilled in the art?

In evaluating whether an AI-generated invention meets applicable patentability requirements, policymakers could consider replacing the person having ordinary skill in the art paradigm with the "AI-aided person having ordinary skill in the art."

However, there could be a risk here that a human examiner, characterised by a much more limited computing capacity than the AI system, might be led to consider an AI-generated invention as always patentable (since, in his or her eyes, the invention is not obvious). It may even be that the examiner is itself AI aided. In this case, if an invention is created by AI, the answer is likely to be that an AI examiner may find everything obvious. This highlights the difficulties and complications that may arise. Clearly, the objectivity of one skilled in the art should always be considered as discussed, whether human or computer-aided.

The paradigm of the "AI-aided person having ordinary skill in the art" seems also to be consistent with EPO guidelines that state that, in conducting his or her analysis, the person having ordinary skill in the art should have access to "the means and capacity for routine work and experimentation which are normal for the field of technology in question."[15] In case of AI-generated inventions such "means" should include both the AI systems and the training data. In fact, it seems inevitable that, in order to evaluate whether an AI-generated invention meets the patentability requirements, the "AI-aided person having ordinary skill in the art" should have free access to all the dataset (including both initial data and training data) that have been used by the AI to generate the invention. However, this may also result in a finding of no inventive step or lack of non-obviousness in the EU and US respectively. The challenge in this analysis is to avoid conflating the concept of an AI-aided invention, that is to say one in which an AI system is used as a tool, and the concept of an AI generated invention where one is examining the notion of the AI system being an inventor. It would seem difficult to accept that the use of a tool in aiding the development of an invention would mean that the tool is an inventor. The foregoing discussion highlights the potential for a lack of clarity when discussing issues surrounding AI because the temptation is to ascribe human qualities to AI systems. In part, this is down to the vernacular using the word "intelligence." It may be better, and more

accurate, to use the phrase "machine learning" although even then the term "learning" implies human qualities. A phrase such as "real-time output feedback responsive adaptive configuration" machine is certainly less accessible than terms such as artificial intelligence or machine learning but may well be more accurate and avoid falling into error when discussing the nature of such machines.

Copyright—deep fakes

The IP issues relating to deep fake infringements were not expressly addressed in the first edition of *Responsible AI,* although deep fakes were addressed briefly in Chapter 1: Ethical Purpose and Societal Benefit. However, this has become a topic of particular interest in relation to the interplay between AI and IP, and is a specific topic addressed in the WIPO Consultation. In this context, we refer to "deep fakes" as being videos (or other digital representations) which have been manipulated by means of deep learning, so as to make the altered video appear to be authentic.

The WIPO Revised Issues Paper[16] states that "technology for deep fakes, or the generation of simulated likenesses of persons and their attributes, such as voice and appearance, exists and is being deployed. Considerable controversy surrounds deep fakes, especially when they have been created without the authorization of a person depicted in the deep fake and when the representation creates actions or attributes views that are not authentic. Some call for the use of deep fake technology to be specifically banned or limited. Others point to the possibility of creating audiovisual works that might allow the deployment of popular or famous performers after their demise in a continuing manner; indeed, it might be possible for a person to authorize such use."

The WIPO Revised Issues Paper[17] queries whether the copyright system should "take cognizance of deep fakes" and queried to whom should the copyright in a deep fake belong, and whether there should be a system of equitable remuneration for those whose likenesses or performances are used in the deep fake. However, more fundamentally, following the responses to the Consultation, the WIPO Revised Issues Paper[18] asks whether copyright is the appropriate vehicle for the regulation of deep fakes and whether deep fakes should benefit from copyright at all, since deep fakes are based on data that may be the subject of copyright.

The questions raised by WIPO in relation to deep fakes, particularly in the WIPO Revised Issues Paper, demonstrate that the appropriate legal position with regard to deep fakes is not straightforward. There may, for example, be a need for different approaches as to ownership and infringement depending on the nature of the deep fake.

Technically the deep fake is a derivative work, which could infringe the copyright in the original work, but which can also be copyrighted in its own right. In this respect, it may be questioned whether the creator of a deep fake could claim copyright ownership of any newly created deep fake and, indeed, who would be the creator for these purposes. However, the same issues as to the need to identify a natural person as the copyright owner will arise as discussed in the first edition of *Responsible AI.*

Another issue is as to whether the use of the original content itself amounts to an infringement of third party copyright and/or image rights (where such rights exist). The questions raised in the WIPO Revised Issues Paper do not focus expressly on infringement. Nevertheless, in this respect, in any claim for infringement, there would be likely to be issues as to the quality and quantity of the original work (or works) used. Moreover, from the infringement perspective, in certain circumstances, deep fakes might count as fair use or fair dealing/parody. Fair use under US law allows for use of copyright works for limited circumstances such as criticism, comment news reporting, teaching, scholarship, and research. Under UK law the fair dealing exception to copyright infringement includes use for criticism or review, reporting current events and parody, caricature or pastiche. Other jurisdictions will have their own exceptions. Moral rights (or the right of integrity), where these exist, may also come into play with regard to deep fakes. It is not unlikely that the manipulation of a work for a deep fake would amount to derogatory treatment of the original work, as a distortion or mutilation, which would entitle the owner of the moral rights to bring an action for an injunction and/or damages and a right to have the infringing copies destroyed. However, it will have to be seen whether actions will be brought to challenge the creation of deep fakes.

Infringement actions can only be brought by the owner of the copyright and/or moral rights, so will not generally give a right of action to any person depicted in the deep fake.

A different approach may be required for ethically wrong and socially harmful deep fakes, particularly those created without the authorisation of the person depicted in the deep fake and when the representation creates actions or attributes views that are not authentic. In these circumstances it is likely that more regulation will be required, including input from social media platforms. (See also Principle 1, Ethical Purpose and Societal Benefit.)

Sui generis IP rights for autonomous AI Invention and data

Although we addressed database rights generally in the first edition of *Responsible AI*, the WIPO Consultation and, now, the WIPO Revised Issues Paper, address the question of data (and new sui generis rights) in some detail. The WIPO Revised Issues Paper considers whether there should be a sui generis system of IP rights for AI-generated inventions in order to adjust innovation incentives for AI,[19] or a separate sui generis system of protection for original literary and artistic works autonomously generated by AI (for example, one offering a reduced term of protection and other limitations, or one treating AI-generated works as performances).[20] However, it also addresses Data in detail[21] with a number of individual specific questions relating to data,[22] focusing on the possible creation of new rights in relation to data (see further below).

So, should we consider further the possibility of creating a sui-generis right for AI-made inventions, remunerating the investment and efforts of the organisation in the development, training and/or implementation of the AI, akin to database protection under EU law—or do existing laws provide enough incentive and reward for investment?

A sui-generis protection of an AI-generated invention, by granting a sui-generis protection to AI-generated systems akin to the database protection under EU laws, has the benefit of granting a minimum of freedom

of access to the underlying information and dataset, since undertakings, in order to obtain the sui-generis protection, would be incentivised to create (at least partly) openly accessible databases of training data.

Access to and the ownership of data are both critical issues for AI development. However, data stored in a data lake in their rawest form, or data generated by the AI as the output of the analysis conducted on the data lake, are not covered by existing sui generis rights relating to databases. Case law shows a very restrictive approach in particular of the Court of Justice of the European Union (CJEU), judges do not easily grant legal protection to databases.

However, the focus has developed in the Revised Issues Paper to address the "general question that arises for the purposes of the present exercise is whether IP policy should go further than the classical system and create new rights in data in response to the new significance that data have assumed as a critical component of AI." In this respect, WIPO is asking nine relevant questions, such as:

- "...what types of data would be the subject of protection? Would any new IP rights be based on the inherent qualities of data (such as its commercial value) or on protection against certain forms of competition or activity in relation to certain classes of data that are deemed to be inappropriate or unfair, or on both?";[23]
- "If new IP rights were to be considered for data, what IP rights would be appropriate, exclusive rights or rights of monetary compensation for use of the data or both?";[24]
- "How would any new IP rights affect or interact with existing policy frameworks in relation to data, such as privacy, security or unfair competition laws or regulations?";[25]
- "If no new IP rights were to be considered for data, should the frameworks of current IP rights, unfair competition laws, trade secrets laws and similar protection regimes, contractual arrangements and technological measures be amended in favour of a stronger economic protection of data?"[26]

Data lakes are the raw material essential for AI training, they contain unstructured data of all forms and sources. They can be privately owned by organisations, or public. Advocating for the recognition of rights in databases and data lakes used by the AI systems (whether by an extension of the existing sui generis right on databases, or a new specific right), may not be needed for private data lakes. Trade secret law is already providing a protection that could be satisfactory for some organisations if the data lakes meet the conditions for trade secret protection: (i) the data lake has to be kept secret, (ii) it must have a commercial value and (iii) must have been subject to reasonable protection measures to keep it secret.

Private databases accessible only by one private organisation, are also protected from unauthorised extractions by many local laws. Such extraction could only result from an unauthorised intrusion on the company's information system, which is sanctioned.

Regarding AI-generated data, it is worth asking if the creation of a new sui generis right, remunerating the investment and efforts of the organisation in the training of the AI, is necessary? New intellectual property rights may be needed in the context of data sharing, as the owner of the intellectual property rights on the data produced by the AI, might be more inclined to share them knowing that the unauthorised extraction and reuse may be penalised.

The concept of "data exclusivity" used in the pharmaceutical business to protect the safety and efficacy data derived from pre-clinical & clinical trials may provide a useful model for at least outline for the protection of datasets relating to AI systems. Indeed, the submission of datasets for establishing the efficacy of an AI system and possibly going to the question of sufficiency of disclosure may require some element of data exclusivity in order to protect the interests of AI and dataset proprietor yet at the same time promote open data sharing.

Trade marks

In the first edition of *Responsible AI*, we addressed issues relating to trade marks and brand protection. Trade marks were not initially addressed in the WIPO Draft Issues Paper, but have now been picked up in the WIPO Revised Issues Paper,[27] on the basis that there may be areas of trade mark law impacted by AI, not least because of the impact of AI on consumer interactions online. WIPO therefore now ask about the impact and any concerns in relation to AI and trade mark law, and, in particular:

- "Do the functions, law and practice of trademarks need to be reconsidered with the increasing use of AI in marketing and the proliferation of AI used by consumers in the context of Internet of Things applications?"[28]
- "Will the use of AI, knowingly or unknowingly, by the consumer for product selection affect brand recognition? Will principles of trademark law, such as distinctiveness, recollection, likelihood of confusion or average consumer need to evolve due to the increasing use of AI? Are these issues for policymakers to consider?"[29]
- "Who is ultimately responsible for AI's actions, in particular when recommendations include infringing products?"[30]
- "Does the use of AI raise unfair competition issues? Is this an issue that the IP system needs to address?"[31]

Trade secrets

In the first edition of *Responsible AI,* we addressed issues relating to trade secrets. Trade secrets were not initially addressed in the WIPO Draft Issues Paper, but have now been picked up in the WIPO Revised Issues Paper,[32] on the basis that they may be used by IP owners where traditional IP rights fail to provide adequate protection. WIPO acknowledge that use of trade secrets will provide an incentive for innovation in AI, but highlight that the lack of disclosure potentially provides a hurdle to open data sharing, and so seek to ascertain whether the current law reflects the right balance in this respect.[33]

The other questions asked are:

- "Should data and AI applications be protectable by trade secrets or is there a social or ethical interest to override existing trade secret protection?"[34]

- "If data and AI applications should not be protected by trade secrets, should any such exception be limited to certain areas of AI, such as data and applications used in judicial decision-making?"[35]
- "If data and AI applications should not be protected by trade secrets, should data and AI applications be protectable by other IP rights?"[36]
- "If data and AI applications should be protected by trade secrets, should there be a mechanism for evidentiary support and practical mechanisms for preserving the confidentiality of trade secrets?"[37]
- "Given the global importance and scope of AI applications, is there a need to harmonize the law of trade secrets at the international level?"[38]
- "Are there seen or unforeseen consequences of trade secrets on bias or trust in AI applications as trade secrets may increase the lack of reproducibility and explainability of AI?"[39]

Conclusion

These developments highlight the extent to which the complex interaction between AI and IP is now being considered on an international basis, whether through the independent but, in fact, consistent decisions of national courts or through the international remit of WIPO. WIPO is hoping to advance a more structured discussion by bringing together member states, academic, scientific and private organisations as well as individuals and other stakeholders to discuss the impact of AI on IP policy. This should help to ensure the necessary consensus in relation to AI and IP rights to allow for the rapid dissemination of new technologies.

Revisions to Principle 8

A few revisions have been made to Principle 8. These have been to identify the need to achieve an appropriate balance between the need to incentivise those developing and using AI technology, and the need to ensure that at least some degree of public benefit is achieved, and the consequent importance of disclosure.

Principle 8

AI and Intellectual Property

Organisations that develop, make available or use AI systems should seek to strike a fair balance between benefiting from adequate protection for the intellectual property rights for both the AI system and the AI output and allowing availability for the wider societal benefit. Governments should investigate how AI systems and AI-created output may be afforded adequate protection whilst also ensuring that the innovation is sufficiently disclosed to promote progress.

1 Supporting incentivisation and protection for innovation

1.1 Innovation is of greatest value when it benefits society. Funding is necessary to develop innovation to a level where it can be disseminated and utilised by society. Those from whom funding is sought require a return on their investment. Consequently, there must be incentivisation and protection for innovation if it is to attract investment and be brought to the greater good of society.

1.2 Organisations must therefore be allowed to protect rights in works resulting from the use of AI, whether AI-created works or AI-enabled works.

1.3 However, care needs to be taken to ensure consistency with the policy objectives of existing intellectual property regimes in order to avoid inconsistencies between respective regimes.

1.4 There should be a balance between the protection of innovation and disclosure of innovation.

2 Protection of IP rights

2.1 The possibility of the creation of works by autonomous AI is likely to require amendments to existing IP laws.

2.2 Organisations that develop, deploy or use AI systems should have the option to take necessary steps to protect the rights for the AI system and in the resulting works. Where appropriate these steps should include asserting or obtaining copyrights, obtaining patents, when applicable, and seeking contractual provisions to allow for protection as trade secrets and/or to allocate the rights appropriately between the parties.

2.3 Nevertheless, the protection of IP rights should not be at the expense of allowing open availability to facilitate development for the wider societal benefit.

3 Development of new IP laws

3.1 Governments should be cautious with revising existing IP laws or seeking to introduce new laws.

3.2 Governments should explore the introduction of appropriate legislation (or the interpretation of existing laws) to clarify IP protection of AI-enabled as well as AI-created output.

3.3 When amending existing or implementing new IP laws, governments should seek adequately to balance the interests of all relevant stakeholders.

3.4 Governments should also explore a consensus in relation to AI and IP rights to promote the unhindered data flows across borders and the rapid dissemination of new technologies and seek to address these issues through an international treaty balancing protection with disclosure.

Endnotes

1 WIPO/IP/AI/2/GE/20/1.

2 https://www.wipo.int/about-ip/en/artificial_intelligence/news/2020/news_0003.html.

3 WIPO/IP/AI/2/GE/20/1 REV.

4 https://www.wipo.int/meetings/en/details.jsp?meeting_id=55309.

5 *Supra* at para 7.

6 The USPTO "Notice to File Missing Parts of Nonprovisional Application" issued on August 8, 2019.

7 In re Application of Application No.: 16/524,350: https://www.uspto.gov/sites/default/files/documents/16524350_22apr2020_3.pdf.

8 https://www.ipo.gov.uk/p-challenge-decision-results/o74119.pdf.

9 https://www.epo.org/news-events/news/2020/20200128.html; https://register.epo.org/application?documentId=E4B63SD62191498&number=EP18275163; https://register.epo.org/application?documentId=E4B63OBI2076498&number=EP18275174.

10 Naruto et al v David Slater: No. 16-15469 (9th Cir. 2018).

11 https://www.federalregister.gov/documents/2019/08/27/2019-18443/request-for-comments-on-patenting-artificial-intelligence-inventions.

12 https://www.uspto.gov/initiatives/artificial-intelligence/notices-artificial-intelligence.

13 *Supra,* Issue 5, para 21.

14 WO2018197835 (A1) https://worldwide.espacenet.com/publicationDetails/biblio?II=0&ND=3&adjacent=true&locale=en_EP&FT=D&date=20181101&CC=WO&NR=2018197835A1&KC=A1#.

15 https://www.epo.org/law-practice/legal-texts/html/guidelines/e/g_vii_3.htm.

16 *Supra* at para 25.

17 *Ibid* at para 26.

18 *Ibid* at para 26(i) and (ii).

19 *Ibid*—Issue 6 at para 22.

20 *Ibid*—Issue 7 at para 23(vii).

21 *Ibid* at paras 28 to 34.

22 *Ibid*—Issue 11 at para 34.

23 *Ibid*—Issue 11 at para 34 (iii).

24 *Ibid*—Issue 11 at para 34 (iv).

25 *Ibid*—Issue 11 at para 34 (vi).

26 *Ibid*—Issue 11 at para 34 (viii).

27 *Ibid*—Issue 13 at paras 36-39.

28 *Ibid* at para 39 (iii).

29 *Ibid* at para 36 (iv).

30 *Ibid* at para 36 (v).

31 *Ibid* at para 36 (vi).

32 *Ibid*—Issue 14 at paras 40-43.

33 *Ibid* at para 42, 43 (i).

34 *Ibid* at para 43 (ii).

35 *Ibid* at para 43 (iii).

36 *Ibid* at para 43 (iv).

37 *Ibid* at para 43 (v).

38 *Ibid* at para 43 (vi).

39 *Ibid* at para 43 (vii).

Responsible AI 2021 Policy Framework

Responsible AI

A GLOBAL POLICY FRAMEWORK

Principle 1
Ethical Purpose and Societal Benefit

Organisations that develop, make available or use AI systems and any national laws or industry standards that govern such use should require the purposes of such implementation to be identified and ensure that such purposes are consistent with the overall ethical purposes of beneficence and non-maleficence, as well as the other principles of the Policy Framework for Responsible AI.

1 Overarching principles

1.1 Organisations that develop, make available or use AI systems should do so in a manner compatible with human agency, human autonomy and the respect for fundamental human rights (including freedom from discrimination).

1.2 Organisations that develop, make available or use AI systems should monitor the implementation of such AI systems and act to mitigate against consequences of such AI systems (whether intended or unintended) that are inconsistent with the ethical purposes of beneficence and non-maleficence, as well as the other principles of the Policy Framework for Responsible AI set out in this framework.

1.3 Organisations that develop, make available or use AI systems should assess the social, political and environmental implications of such development, deployment and use in the context of a structured Responsible AI Impact Assessment that assesses risk of harm and, as the case may be, proposes mitigation strategies in relation to such risks.

2 Human Agency and Autonomy

2.1 Organisations that develop, make available or use AI systems that surveil human behavior shall put in place appropriate safeguards to promote the right to be let alone (the right not to be subject to arbitrary interference with his privacy, family, home or correspondence), informed human agency and autonomy and to avoid destructive self-censorship, loss of individuality and identity, loss of freedom of expression and the loss of human ability to think freely and independently. Such safeguards shall include conducting a responsible AI ethical risk assessment of the technology as part of an accountable governance process prior to deployment of the AI System and ensuring that any such deployment is consistent with respect for other principles of the Policy Framework for Responsible AI such as Transparency and Explainability, Fairness and Non-Discrimination, and Privacy

2.2 Organisations that develop, make available or use AI systems that surveil human behavior using sensitive personal data (such as data collected in non-public spaces such as the home), facial-recognition data or biometric data shall apply the Transparency and Privacy principles with particular rigour, including as regards the reasonable purpose, limited collection, limited use, limited disclosure and limited retention principles, as well as by providing full transparency as to whether and when a device's voice, movement or image surveillance features have been activated. Sensitive personal data such as biometric data and genetic data collected locally by IoT devices (such as fitness monitors and smart phones) and natural language, movement and image data collected by "always on" IoT devices (such as personal assistants and smart home devices) shall, to the great-

est extent possible, securely store such data, in encrypted format, only locally on the device in a manner that allows for the maximal level of autonomy and control over the data by the individual(s) to whom it relates.

2.3 Organisations that develop, make available or use AI systems that predict and influence human behavior shall put in place appropriate safeguards to promote informed human agency and autonomy and to avoid destructive psychological and behavioural manipulation, addiction, dependency and attention deficit. Such safeguards shall include conducting a responsible AI ethical risk assessment of the technology as part of an accountable governance process prior to deployment of the AI System and ensuring that any such deployment is consistent with respect for other principles of the Policy Framework for Responsible AI such as Transparency and Explainability, Fairness and Non-Discrimination, and Privacy.

3 Work and automation

3.1 Organisations that implement AI systems in the workplace should provide opportunities for affected employees to participate in the decision-making process related to such implementation.

3.2 Consideration should be given as to whether it is achievable from a technological perspective to ensure that all possible occurrences should be pre-decided within an AI system to ensure consistent behaviour. If this is not practicable, organisations developing, deploying or using AI systems should consider at the very least the extent to which they are able to confine the decision outcomes of an AI system to a reasonable, non-aberrant range of responses, taking into account the wider context, the impact of the decision and the moral appropriateness of "weighing the unweighable" such as life vs. life.

3.3 Organisations that develop, make available or use AI systems that have an impact on employment should conduct a Responsible AI Impact Assessment to determine the net effects of such implementation.

3.4 Organisations that develop, make available or use AI systems that surveil or influence employee behavior in the workplace shall put in place appropriate safeguards to promote the informed human agency, autonomy and dignity of employees and to avoid inappropriate or destructive impacts on the emotional or psychological health of employees (monotony of tasks, excessive surveillance, gaming of behavior, continuous exposure to horrific content). Such safeguards shall include conducting a responsible AI ethical risk assessment of the technology as part of an accountable governance process prior to deployment of the AI System and ensuring that any such deployment is consistent with respect for other principles of the Policy Framework for Responsible AI such as Transparency and Explainability, Fairness and Non-Discrimination, and Privacy.

3.5 Governments should closely monitor the progress of AI-driven automation in order to identify the sectors of their economy where human workers are the most affected. Governments should actively solicit and monitor industry, employee and other stakeholder data and commentary regarding the impact of AI systems on the workplace and should develop an open forum for sharing experience and best practices.

3.6 Governments should promote educational policies that equip all children with the skills, knowledge and qualities required by the new economy and that promote life-long learning.

3.7 Governments should encourage the creation of opportunities for adults to learn new useful skills, especially for those displaced by automation.

3.8 Governments should study the viability and advisability of new social welfare and benefit systems to help reduce, where warranted, socio-economic inequality caused by the introduction of AI systems and robotic automation.

4 Environmental impact

4.1 Organisations that develop, make available or use AI systems should assess the overall environmental impact of such AI systems, throughout their implementation, including consumption of resources, energy costs of data storage and processing and the net energy efficiencies or environmental benefits that they may produce. Organisations should seek to promote and implement uses of AI systems with a view to achieving overall carbon neutrality or carbon reduction.

4.2 Governments are encouraged to adjust regulatory regimes and/or promote industry self-regulatory regimes concerning market-entry and/or adoption of AI systems in a way that the possible exposure (in terms of 'opportunities vs. risks') that may result from the public operation of such AI systems is reasonably reflected. Special regimes for intermediary and limited admissions to enable testing and refining of the operation of the AI system can help to expedite the completion of the AI system and improve its safety and reliability.

4.3 In order to ensure and maintain public trust in final human control, governments should consider implementing rules that ensure comprehensive and transparent investigation of such adverse and unanticipated outcomes of AI systems that have occurred through their usage, in particular if these outcomes have lethal or injurious consequences for the humans using such systems. Such investigations should be used for considering adjusting the regulatory framework for AI systems, in particular to develop, where practicable and achievable, a more rounded understanding of how and when such systems should gracefully handover to their human operators in a failure scenario.

4.4 AI has a particular potential to reduce environmentally harmful resource waste and inefficiencies. AI research regarding these objectives should be encouraged. In order to do so, policies must be put in place to ensure the relevant data is accessible and usable in a manner consistent with respect for other principles of the Policy Framework for Responsible AI such as Fairness and Non-Discrimination, Open Data and Fair Competition and Privacy.

5 Weaponised AI

5.1 The use of lethal autonomous weapons systems (LAWS) should respect the principles and standards of and be consistent with international humanitarian law on the use of weapons and wider international human rights law.

5.2 Governments should implement multilateral mechanisms to define, implement and monitor compliance with international agreements regarding the ethical development, use and commerce of LAWS.

5.3 Governments and organisations should refrain from developing, selling or using lethal autonomous weapon systems (LAWS) able to select and engage targets without human control and oversight in all contexts.

5.4 Organisations that develop, make available or use AI systems should inform their employees when they are assigned to projects relating to LAWS.

6 The weaponisation of false or misleading information

6.1 Organisations that develop, make available or use AI systems to filter or promote informational content on internet platforms that is shared or seen by their users should take reasonable measures, consistent with applicable law, to

minimise the spread of false or misleading information where there is a material risk that such false or misleading information might lead to significant harm to individuals, groups or democratic institutions.

6.2 AI has the potential to assist in efficiently and pro-actively identifying (and, where appropriate, suppressing) unlawful content such as hate speech or weaponised false or misleading information. AI research into means of accomplishing these objectives in a manner consistent with freedom of expression should be encouraged.

6.3 Organisations that develop, make available or use AI systems on platforms to filter or promote informational content that is shared or seen by their users should provide a mechanism by which users can flag potentially harmful content in a timely manner.

6.4 Organisations that develop, make available or use AI systems on platforms to filter or promote informational content that is shared or seen by their users should provide a mechanism by which content providers can challenge the removal of such content by such organisations from their network or platform in a timely manner.

6.5 Governments should provide clear guidelines to help organisations that develop, make available or use AI systems on platforms identify prohibited content that respect both the rights to dignity and equality and the right to freedom of expression.

6.6 Courts should remain the ultimate arbiters of lawful content.

Principle 2
Accountability

Organisations that develop, make available or use AI systems ought to be accountable for the consequences of their actions and shall designate an individual or individuals who are accountable for the organisation's compliance with the principles of the Policy Framework for Responsible AI or other adopted principles (including analogous principles that may be developed for a specific industry) with the objective of keeping humans behind the machines and AI Human centric.

1 Accountability

1.1. The identity of the individual(s) designated by the organisation to oversee the organisation's compliance with the principles shall be made known upon request.

1.2. Organisations that develop, make available deploy or use AI systems shall use human oversight to carry out determination of the situations in which to carry out delegation to AI decision-making, while ensuring that such use is to accomplish human-chosen objectives. Human oversight can be achieved through three mechanisms, i.e. human-in-the-loop (where humans retain full control to intervene in every decision-making cycle), human-on-the-loop (where humans can intervene during the design cycle of the system and may carry out monitoring) and human-in-command (where humans can oversee the overall activity of the AI system and decide the situations and manner in which it may be used).

1.3. Organisations that develop, make available deploy or use AI systems shall implement policies and practices to give effect to the principles of the Policy Framework for Responsible AI or other adopted principles (including analogous principles that may be developed for a specific industry), including:

i. establishing processes to determine whether, when and how to implement a "Responsible AI Impact Assessment" process;

ii. establishing and implementing "Responsible AI by Design" principles;

iii. establishing procedures to receive and respond to complaints and inquiries;

iv. training staff and communicating to staff information about the organisation's principles, policies and practices; and

v. developing information to explain the organisation's principles, policies and procedures.

2 Government

2.1. Governments should seek to work collaboratively and in a coordinated manner across the international landscape to apply the principles of this Policy Framework for Responsible AI or other analogous internationally recognised principles to ensure consistency of approach and application when holding AI systems to account.

2.2. Governments that assess the potential for "accountability gaps" in existing legal and regulatory frameworks applicable to AI systems

should adopt a balanced approach that encourages innovation while mitigating against the risk of significant individual or societal harm.

2.3. Any such legal and regulatory frameworks should promote the eight principles of the Policy Framework for Responsible AI or encompass similar considerations and consider appropriate legal and regulatory enforcement and redress mechanisms.

2.4. Governments should not grant distinct legal personality to AI systems, as doing so would undermine the fundamental principle that humans should ultimately remain accountable for the acts and omissions of AI systems.

2.5. Governments should be transparent and put appropriate human oversight mechanisms in place when utilising AI systems for products or services which are in the public interest, and ensure that the objective and outcomes of such AI Systems are understood by its subjects or citizens.

3 Contextual approach

3.1. The intensity of the accountability obligation will vary according to the degree of autonomy and criticality of the AI system and its potential to cause individual or societal harm. The greater the level of autonomy of the AI system and the greater the criticality of the outcomes that it may produce, the higher the degree of accountability that will apply to the organisation that develops, deploys or uses the AI system ("High Risk AI").

3.2. Where an AI system is deemed to be High Risk AI, a Responsible AI Impact Assessment ("RAIIA") should be conducted and clearly identify the accountable person(s).

Principle 3
Transparency and Explainability

Organisations that develop, make available or use AI systems, and any national laws or industry standards that govern such use, shall ensure that such use is transparent and that the decision outcomes of the AI system are explainable.

1 Purpose

1.1 The Transparency and Explainability principle aims to promote and maintain public trust in AI systems by requiring organisations that develop, make available and use AI systems to provide sufficient information to demonstrate whether decisions made by the AI systems are fair and impartial, support human agency and human autonomy and establish meaningful responsibility and accountability of an AI system's developers and users.

1.2 The Transparency and Explainability principle supports the Ethical Purpose and Societal Benefit principle, the Accountability principle, the Fairness and Non-Discrimination principle, the Safety and Reliability principle and the Privacy principle.

2 Transparency

2.1 Organisations that make available or use an AI system in decision-making processes which produce legal effects concerning an individual or similarly significantly affects an individual shall make readily available meaningful information regarding: (a) the fact that an AI system is being used in a decision-making process; (b) the intended purpose(s); (c) the types of data sets that are used and generated by the AI system; and (d) whether and to what extent the decision-making process may include human participation.

2.2 The information set forth in Section 2.1 should be made readily available to the affected individual before such automated decision-making process occurs in order to provide the individual with an opportunity to assess whether or not to seek a human-centric alternative decision-making process.

3 Explainability

3.1 Organisations that make available or use an AI system in decision-making processes which produce legal effects concerning an individual or similarly significantly affects an individual shall make readily available to such individuals information in objectively clear terms that explains how a decision/outcome was reached, with, at a minimum: a) the information set forth in Section 2.1 above; b) information that offers meaningful interpretability of the algorithmic logic of the AI system; c) meaningful information to understand the decision/outcome; and d) information regarding how the individual may contest the decision or outcome.

3.2 The information set forth in Section 3.1 should be made readily available to an affected individual promptly after such automated decision-making process occurs in order to provide the affected individual with an opportunity to assess whether or not to challenge the decision or outcome.

4 Gradual and contextual approach

4.1 The intensity of the transparency and explainability obligations will depend on a variety of factors, including the nature of the data involved, lack of human participation in the decision-making, the result of the decision and its consequences for the affected individual.

4.2 Ultimately, transparency and explainability must balance the rights, interests and reasonable expectations of the person subject to the decision with the legitimate interests of the organisation making the decision and considerations of overall societal benefit.

4.3 The intensity of the transparency and explainability obligations will generally be higher where the AI system is made available or used in relation to lay persons who are unlikely to understand the technology rather than with an expert whose understanding of the system may be more easily established. Moreover, the intensity of the transparency and explainability obligations will generally be higher where an AI system is used by a public sector organization in the context of enforcing legal obligations rather than by a private sector organisation in the context of offering services.

4.4 The intensity of the transparency and explainability obligations will generally be higher where sensitive personal data is used or where the outcome of the decision will have a material impact on the affected individual's legal or human rights or similarly significantly affects an individual. The intensity of these obligations will generally be lower where non-sensitive personal data or de-personalised data is used or where the impacts on the affected individual's legal or human rights are relatively inconsequential.

4.5 In situations giving rise to high intensity transparency and explainability obligations, organisations that make available or use an AI system in decision-making processes affecting individual rights should, in addition to the information set forth in Sections 2.1 and 3.1 above, make readily available to such individuals meaningful information regarding: a) the traceability and auditability of the algorithmic logic of the AI system, and b) the testing methods used to promote the principles within this policy framework.

5 Transparency and explainability by design

5.1 Organisations that develop AI systems should ensure that the system architecture, algorithmic logic, data sets, testing methods, and all related development and operational policies and procedures serve to incorporate and embed transparency and explainability by design in accordance with national laws and consistent with relevant industry standards. In so far as is reasonably practicable, such systems should aim to be designed from the outset and maintained to promote meaningful transparency and explainability that complements the intended purpose(s) of the AI system.

4.2 The design and development methodologies adopted in Section 5.1 should have the flexibility to embrace evolving industry standards, providing ongoing iterative improvements in transparency and explainability in parallel with advancement in the state of the art during the lifecycle of the AI system.

4.3 Since embedding transparency and explainability into AI system design requires extensive planning and multi-disciplinary expertise, organisations should develop frameworks to assist programmers and developers to design and develop AI systems that possess the desired values and to help reconcile the tensions that exist between accuracy, cost and explainability.

6 Technological neutrality

6.1 The use of an AI system by an organisation does not increase or reduce the procedural and substantive requirements that would otherwise apply if the decision-making process were controlled by a human.

Principle 4
Fairness and Non-Discrimination

Organisations that develop, make available or use AI systems and any national laws that regulate such use shall ensure the non-discrimination of AI outcomes, and shall promote appropriate and effective measures to safeguard fairness in AI use.

1 Awareness and education

1.1 Awareness and education on the possibilities and limits of AI systems is a prerequisite to achieving fairer outcomes.

1.2 Organisations that develop, make available or use AI systems should take steps to ensure that users are aware that AI systems reflect the goals, knowledge and experience of their creators, as well as the limitations of the data sets that are used to train them.

2 Technology and fairness

2.1 Carefully designed AI systems offer the possibility of more consistently fair and non-discriminatory outcomes than are achievable in systems that rely on human decision-making.

2.2 Decisions based on AI systems should be fair and non-discriminatory, judged against the same standards as decision-making processes conducted entirely by humans.

2.3 The use of AI systems by organisations that develop, make available or use AI systems and Governments should not serve to exempt or attenuate the need for fairness, although it may mean refocusing applicable concepts, standards and rules to accommodate AI.

2.4 Users of AI systems and persons subject to their decisions must have an effective way to seek remedy in discriminatory or unfair situations generated by biased or erroneous AI systems, whether used by organisations that develop, make available or use AI systems or governments, and to obtain redress for any harm. Taking into consideration the societal impacts of unfair AI, collective remedies could be a useful tool to address bias or unfairness.

3 Development and monitoring of AI systems

3.1 AI development should be designed to prioritise fairness and non-discrimination. This would involve addressing algorithms and data bias from an early stage and continuously throughout the entire lifecycle of the AI system with a view to ensuring fairness and non-discrimination.

3.2. Before making available or using an AI system, organisations should systematically assess the expected performance of the AI system with respect to potentially unlawful or unfair discrimination as compared to the performance of the processes currently in use.

3.3. Organisations that develop, make available or use AI systems should remain vigilant to the dangers posed by bias. This could be achieved by establishing ethics boards and codes of conduct, and by adopting industry-wide standards and internationally recognised quality seals.

3.4. AI systems with an important social impact could require independent reviewing and testing on a periodic basis.

3.5. In the development and monitoring of AI systems, particular attention should be paid to disadvantaged groups which may be inadequately or unfairly represented in the training data.

4 A comprehensive approach to fairness

4.1 AI systems can perpetuate and exacerbate bias, and have a broad social and economic impact in society. Addressing non-discrimination and fairness in AI use requires a holistic approach. In particular, it requires:

i. the close engagement of technical experts from AI-related fields with statisticians and researchers from the social sciences; and

ii. a combined engagement between governments, organisations that develop, make available or use AI systems and the public at large.

4.2 The Fairness and Non-Discrimination Principle is supported by the Transparency and Accountability Principles. Effective fairness in use of AI systems requires the implementation of measures in connection with both these Principles.

Principle 5
Safety and Reliability

Organisations that develop, make available or use AI systems and any national laws that regulate such use shall adopt design regimes and standards ensuring high safety and reliability of AI systems on one hand while limiting the exposure of developers and deployers on the other hand.

1 Require and/or define explicit ethical and moral principles underpinning the AI system

1.1 Governments and organisations developing, making available or using AI systems should define the relevant set of ethical and moral principles underpinning the AI system to be developed, deployed or used taking into account all relevant circumstances. A system designed to autonomously make decisions will only be acceptable if it operates on the basis of clearly defined principles and within boundaries limiting its decision-making powers.

1.2 Governments and organisations developing, making available or using AI systems should validate the underpinning ethical and moral principles as defined periodically to ensure ongoing accurateness.

2 Standardisation of behaviour

2.1 Governments and organisations developing, making available or using AI systems should recall that ethical and moral principles are not globally uniform but may be impacted e.g. by geographical, religious or social considerations and traditions. To be accepted, AI systems might have to be adjustable in order to meet the local standards in which they will be used.

2.2 Consider whether all possible occurrences should be pre-decided in a way to ensure the consistent behaviour of the AI system, the impact of this on the aggregation of consequences and the moral appropriateness of "weighing the unweighable" such as life vs. life.

3 Ensuring safety, reliability and trust

3.1 Governments should require and organisations should test AI systems thoroughly to ensure that they reliably and robustly adhere, in operation, to the underpinning ethical and moral principles and have been trained with data which are curated and are as 'error-free', 'bias-free' as practicable, given the circumstances. This includes requirements on procedural transparency and technical transparency of the development process of the AI system and the data uses in that respect, as well as the explainability of the decision-making process an AI system will apply when in operation.

3.2 Governments are encouraged to adjust regulatory regimes and/or promote industry self-regulatory regimes for allowing market-entry of AI systems in order to reasonably reflect the positive exposure that may result from the public operation of such AI systems. Special regimes for intermediary and limited admissions to enable testing and refining of the operation of the AI system can help to expedite the completion of the AI system and improve its safety and reliability.

3.3 In order to ensure and maintain public trust in final human control, governments should consider implementing rules that ensure com-

prehensive and transparent investigation of such adverse and unanticipated outcomes of AI systems that have occurred through their usage, in particular if these outcomes have lethal or injurious consequences for the humans using such systems. Such investigations should be used for considering adjusting the regulatory framework for AI systems; in particular to develop a more rounded understanding of how such systems should gracefully handover to their human operators.

4 Facilitating technological progress at reasonable risks

4.1 Governments are encouraged to consider whether existing legal frameworks such as product liability require adjustment in light of the unique characteristics of AI systems.

4.2 As AI systems might be partially autonomous, organisations developing, deploying or using such systems should pursue continuous monitoring of systems deployed and/or used, allowing human operators to interrupt unanticipated alterations.

4.3 Governments should support and participate in international co-ordination (through bodies such as the International Organisation for Standardisation (ISO) and the International Electrotechnical Commission (IEC)) to develop international standards for the development and deployment of safe and reliable AI systems. Governments are further encouraged to contemplate requirements on continuous monitoring with human oversight as part of their regime balancing encouragement of progress vs. risk avoidance.

Principle 6
Open Data and Fair Competition

Organisations that develop, make available or use AI systems and any national laws that regulate such use shall, without prejudice to normal rules of intellectual property and privacy:

(a) foster open access to, and the portability of, datasets (where privately held), especially where such datasets are deemed significant and important or advance the "state of the art" in the development of AI systems;

(b) ensure that data held by public sector bodies are, in so far as is reasonably practicable, portable, accessible and open; and

(c) encourage open source frameworks and software for AI systems which could similarly be regarded as significant and important and advance the "state of the art."

AI systems must be developed and made available on a "compliance by design" basis in relation to competition/antitrust law.

1 Supporting effective competition in relation to AI systems

1.1 Governments should support and participate in international co-ordination (through bodies such as the OECD and the International Competition Network) to develop best practices and rigorous analysis in understanding the competitive impact of dataset control and AI systems on economic markets.

1.2 Governments should undertake regular reviews to ensure that competition law frameworks and the enforcement tools available to the relevant enforcement authorities are sufficient and effective to ensure sufficient access to necessary inputs, and adequate choice, vibrant rivalry, creative innovation and high quality of output in the development and deployment of AI systems, to the ultimate benefit of consumers.

2 Open data

2.1 Governments should foster and facilitate national infrastructures necessary to promote the portability of and open access to, datasets, especially those that are significant and important, to all elements of society having a vested interest in access to such datasets for research and/or non-commercial use to further advance the "state of the art" in relation to such technology and to ensure the efficacy of existing AI systems. In this regard, governments should give serious consideration to two-tier access models which would allow for free access for academic and research purposes, and paid-for access for commercialised purposes.

2.2 Governments should support open data initiatives in the public or private sector with guidance and research to share wide understanding of the advantages to be gained from open access data, the structures through which datasets can be shared and exchanged, and the processes by which data can be made porta-

ble and suitable for open access (including API standardisation, pseudonymisation, aggregation or other curation, where necessary).

2.3 Governments should ensure that the data held by public sector bodies are accessible and open, where possible and where this does not conflict with a public sector mandate to recover taxpayer investment in the collection and curation of such data. Private sector bodies such as industry organisations and trade associations should similarly support and promote open data within their industry sector, making their own datasets open, where possible. The degree of relative influence that private sector organisations have on applicable markets should be assessed on a continuous basis by regulators.

2.4 Organisations that develop, make available or use datasets, especially those which could be regarded as significant or important or which could be regarded as advancing the "state of the art" are similarly encouraged to open up access to, and/or license, such datasets, where possible via chaperoned mechanisms such as Data Trusts.

2.5 Any sharing or licensing of data should be to an extent which is reasonable in the circumstances and must be in compliance with legal, regulatory, contractual and any other obligations or requirements in relation to the data concerned (including privacy, security, freedom of information and other confidentiality considerations). In addition, all stakeholders involved in such sharing or licensing should be very clearly identified in terms of legal roles, duties and responsibilities.

3 Open source AI systems

3.1 Organisations that develop AI systems are normally entitled to commercialise such systems as they wish. However, governments should at a minimum advocate accessibility through open source or other similar licensing arrangements to those innovative AI systems which may be of particular societal benefit or advance the "state of the art" in the field via, for example, targeted incentive schemes.

3.2 Organisations that elect not to release their AI systems as open source software are encouraged nevertheless to license the System on a commercial basis.

3.3 To the extent that an AI system can be subdivided into various constituent parts with general utility and application in other AI use-cases, organisations that elect not to license the AI system as a whole (whether on an open source or commercial basis) are encouraged to license as many of such re-usable components as is possible.

4 Compliance by design with competition/antitrust laws

4.1 Organisations that develop, deploy or use AI systems should design, develop and deploy AI systems in a "compliance by design" manner which ensures consistency with the overarching ethos of subsisting competition/antitrust regimes to promote free and vibrant competition amongst corporate enterprises to the ultimate benefit of consumers.

Principle 7
Privacy

Organisations that develop, make available or use AI systems and any national laws that regulate such use shall endeavour to ensure that AI systems are compliant with privacy norms and regulations, taking into account the unique characteristics of AI systems, and the evolution of standards on privacy.

1 Finding a balance

1.1 There is an inherent and developing conflict between the increasing use of AI systems to process private data, especially personal data; and the increasing regulatory protection afforded internationally to personal and other private data This protection typically applies principles of purpose limitation, data minimisation and storage limitation.

1.2 Governments that regulate the privacy implications of AI systems should do so in a manner that acknowledges the specific characteristics of AI and that does not unduly stifle AI innovation.

1.3 However, governments should foster the privacy principles, in particular of purpose limitation, for personal data within the use of AI systems.

1.3 Organisations that develop, make available and use AI systems should analyse and constantly check their current processes to identify whether they need be updated or amended in any way to ensure that the respect for privacy is given as a central consideration. This includes consideration as to whether and to want extent AI systems actually require the processing of personal (as opposed to, e.g. anonymous) data.

2 The operational challenges ahead for AI users

2.1 AI systems create challenges specifically in relation to the practicalities of meeting of requirements under a number of national legislative regimes, such as in relation to consent and anonymisation of data. Likewise AI systems create challenges as to data subject rights and legal certainty for all parties involved. Accordingly, organisations that develop, deploy or use AI systems and any national laws that regulate such use, shall make provision for alternative lawful bases for the collection and processing of personal data by AI systems, such as a rightful use of the input and output data.

2.2 Organisations that develop, deploy or use AI systems should identify the level of responsibility when they use input or output data for AI systems (e.g. to avoid unlawful discrimination). Organisations should then consider the resulting consequences and obligations, including implementing operational safeguards to protect privacy such as privacy by design principles that are specifically tailored to the specific features of deployed AI systems.

2.3 Organisations that develop, deploy and use AI systems should appoint an AI Ethics Officer, in a role similar to Data Protection Officers under the GDPR, but with specific remit to consider the ethics and regulatory compliance of their use of AI.

3 AI as a tool to support privacy

3.1 Although there are challenges from a privacy perspective from the use of AI, in turn the advent of AI technologies could also be used to help organisations comply with privacy obligations.

Principle 8
AI and Intellectual Property

Organisations that develop, make available or use AI systems should seek to strike a fair balance between benefiting from adequate protection for the intellectual property rights for both the AI system and the AI output and allowing availability for the wider societal benefit. Governments should investigate how AI systems and AI-created output may be afforded adequate protection whilst also ensuring that the innovation is sufficiently disclosed to promote progress.

1 Supporting incentivisation and protection for innovation

1.1 Innovation is of greatest value when it benefits society. Funding is necessary to develop innovation to a level where it can be disseminated and utilised by society. Those from whom funding is sought require a return on their investment. Consequently, there must be incentivisation and protection for innovation if it is to attract investment and be brought to the greater good of society.

1.2 Organisations must therefore be allowed to protect rights in works resulting from the use of AI, whether AI-created works or AI-enabled works.

1.3 However, care needs to be taken to ensure consistency with the policy objectives of existing intellectual property regimes in order to avoid inconsistencies between respective regimes.

1.4 There should be a balance between the protection of innovation and disclosure of innovation.

2 Protection of IP rights

2.1 The possibility of the creation of works by autonomous AI is likely to require amendments to existing IP laws.

2.2 Organisations that develop, deploy or use AI systems should have the option to take necessary steps to protect the rights for the AI system and in the resulting works. Where appropriate these steps should include asserting or obtaining copyrights, obtaining patents, when applicable, and seeking contractual provisions to allow for protection as trade secrets and/or to allocate the rights appropriately between the parties.

2.3 Nevertheless, the protection of IP rights should not be at the expense of allowing open availability to facilitate development for the wider societal benefit.

3 Development of new IP laws

3.1 Governments should be cautious with revising existing IP laws or seeking to introduce new laws.

3.2 Governments should explore the introduction of appropriate legislation (or the interpretation of existing laws) to clarify IP protection of AI-enabled as well as AI-created output.

3.3 When amending existing or implementing new IP laws, governments should seek adequately to balance the interests of all relevant stakeholders.

3.4 Governments should also explore a consensus in relation to AI and IP rights to promote the unhindered data flows across borders and the rapid dissemination of new technologies and seek to address these issues through an international treaty balancing protection with disclosure.

Responsible AI Impact Assessment Tool (RAIIA)

Responsible AI

A GLOBAL POLICY FRAMEWORK

Responsible AI Impact Assessment ("RAIIA") Template

Version 1.1 [3rd Jan 2021]

Company: _____

Date: _____

Disclaimer

This template is provided as is without any warranties of any kind. It should only be used as a guide whilst evaluating an AI System. Adjust it as necessary to fit your needs.

How to Use This Template

Index

Document Name(s) (specify):

_____

_____

_____

Assessment Instructions

RAIIA Steps:

a. Fill in the essential information about the AI System and the Project in the Project Summary Section.
b. Determine if a RAIIA is necessary by evaluating key risks factors as per the Key Factors for RAIIA.
c. For each Principle, answer each questions in as much detail as possible, determining how the AI System will impact or address the risk factor.
d. Determine a risk rating for each risk factor of every principle.
e. Based on the risk rating, determine tailored mitigation measures to reduce the initial risk rating.
f. Consider the impact of the mitigation measures on the risk factor and revise the initial risk rating.
g. Review periodically.

Assessment of Risks

Risks Level	Description of Risks
QA	
0	Zero
1	Very Low
2	Low
3	Medium
4	High
5	Very high

Glossary

AI System	Solution or product to be developed or deployed with data-driven, predictive functionality based upon any artificial intelligence or machine learning capability
Organization	Business or entity with the goal to implement an AI System and that will conduct the RAIIA and ultimately drive the Project
Project	The application use case which will be implemented, resolved or managed using the AI System

1. Project Summary

This Project Summary section is used to provide a high level summary of the Project, the AI System and the business context in which it is implemented.

References to "AI System" means an AI software solution, or product to be developed or deployed as part of the Project.

Background

Project Name	
Business Segment	
Line of Business Name	
RAIIA Evaluation Date	
Project Start Date	
AI System Launch Date	
Region	
Person responsible for the RAIIA	
Status of Assessment	

Summary

High Level Technical and Functional overview	
Business driver and context	
Data sources and data sets	
→ Internal	
→ External	
Summary of potential risks	
→ Legal	
→ Ethical	
→ Environmental	
→ Reputational	
External Related Documents	
Governance model	
Project team	

2. Key Factors for Conducting a RAIIA

The following questions should be answered to assess whether an RAIIA is necessary or appropriate for the AI System.

This list is not exhaustive and may need to be tailored to the specific context of the Organization and AI System, and adjusted to reflect evolving standards and applicable law.

Risk factors should be evaluated based on a **0 to 5** scale (zero risk to very high risk).

A holistic and contextual approach is recommended. Such an approach should consider the factors in relation to one another.

Supplemental content (including documents) for your particular use case should be referenced and listed.

	Factors to Evaluate Need for RAIIA	Answers	Predicted Risk Rating
Context	1. Describe the context in which the AI System will be used or deployed.		
	2. Will the use of the AI System be citizen-facing?		
	3. What is the market, industry or sector targeted?		
Laws and regulations	4. Do the jurisdiction(s) in which the AI Solution will be deployed have data protection laws or regulation that are applicable to its use?		
	5. Does the jurisdiction(s) in which the Project will take place abide by rule of law principles?		
	6. Does this jurisdiction have antidiscrimination laws?		
	7. What are the main regulatory requirements relevant to the use and deployment of the AI System within the targeted market, industry or sector?		
	8. Will the AI System be used across legal jurisdiction borders (whether they be across federal states or national borders)?		
	9. What are the main ethical concerns relevant to the use and deployment of the AI System for the targeted market, industry or sector?		
Stakeholders	10. Who will be the main stakeholders affected by the AI System?		
	11. Who are the expected contributing third parties?		
	12. What individual rights and interests will be at stake as a consequence of the use of the AI System?		
	13. Are those rights fundamental or human rights?		

	Factors to Evaluate Need for RAIIA	Answers	Predicted Risk Rating
Human Oversight	14. Will the AI System make or participate in making decisions with material impacts on individuals or society?		
	15. What is the expected degree of autonomy of the AI System? Will, for instance, human operators or decision-makers have oversight on individual AI decisions, if any?		
	16. How frequently will there be human oversight over the operation of the AI System?		
	17. What measures would be taken to avoid automation bias or anchoring to the AI System?		
	18. What will be the Organisation's degree of control and responsibility over the finalized AI System?		
Data and Privacy	19. What is the type and origin of the data that will be used to train the AI System?		
	20. Will the training data include personal information?		
	21. If personal information are used in the context of the AI System, who are the data subjects?		
	22. What is the level of sensitivity of the data in term of privacy?		
Human-understandable AI	23. What are the technical characteristics of the AI System that could influence the explainability and auditability of the algorithm?		
	24. Can the results of the AI System be explained in humanly understandable terms?		
			Total

Total Predicted Risk Rating	Risk Level	Comments
Below 11	Very low	RAIIA not necessary
Between 12 and 22	Low	RAIIA not necessary
Between 23 and 33	Medium	RAIIA recommended
Between 34 and 44	High	RAIIA highly recommended
Above 45	Very high	RAIIA Necessary

3. Main Assessment

Principle 1: Ethical Purpose and Societal Benefit

Organisations that develop, make available or use AI and any national laws that regulate such use should require the purposes of such implementation to be identified and ensure that such purposes are consistent with the overall ethical purposes of beneficence and non-maleficence, as well as the other principles, in particular those of the Policy Framework for Responsible AI.

Overview of Principle 1

Organisations that develop, make available or use AI Systems should:

- do so in a manner compatible with human agency, human autonomy and the respect for fundamental human rights (including freedom from discrimination);
- monitor the implementation of such AI Systems and to act to mitigate against consequences of such AI Systems (whether intended or unintended) that are inconsistent with the ethical purposes of beneficence and non-maleficence;
- assess the social, political and environmental implications of such development, deployment and use in the context of a structured Responsible AI Impact Assessment that assumes risk of harm and, as the case may be, proposes mitigation strategies in relation to such risks.

Risk Factors	Whether/How the Solution Addresses the Factors	Risk Rating	Mitigation Measures	Revised Risk Rating
1. Is the AI System consistent with the ethical principles, values, standards, policies and/or code of conduct of the Organisation?				
2. Are there any potential reputational and material risks attached to the AI System for the Organisation?				
3. Is there a risk that use of the AI System will violate any fundamental human rights (such as rights of freedom, free expression, non-discrimination)?				
4. Does the AI System raise risks to human agency (such as self-determination, choice, free will, unfettered decision making, and the ability to self-regulate one's own affairs) in respect of the intended end user audience or other ecosystem stakeholders?				
5. Does the AI System raise risks to human autonomy (such as freedom of movement and travel; data portability) in respect of the intended end user audience or other ecosystem stakeholders?				
6. Is there a risk(s) that the AI System could generate confusion as to whether or not the user is interacting with a human or an AI System?				
7. Does the AI System involve surreptitious surveillance or excessive surveillance that might impose a danger to human agency and autonomy (such as encouraging self-censorship or limiting freedom of expression or assembly)?				

Risk Factors	Whether/How the Solution Addresses the Factors	Risk Rating	Mitigation Measures	Revised Risk Rating
8. Does the AI System promote over-reliance, dependency, addiction or attention deficit?				
9. Does the AI System raise risks of psychological and behavioural manipulation, coercion or excessive nudging?				
10. Does the AI System hinder the user's ability to make informed decisions?				
11. Contrary to Q9, does the AI System empower the user?				
12. Is there a risk that the AI System will promote the spread of false or misleading information?				
13. Is there a risk that the AI System will promote the spread of hate speech, unlawful content or content which is potentially dangerous (physically, psychologically, or emotionally) to the end recipients/viewers of the content?				
14. Are there employment-related risks associated with the AI System (such as material job loss or functionality that might detrimentally affect the quality of work experience)?				
15. Contrary to Q14, does the AI System serve primarily to empower workers (by providing them with effective tools, skills or knowledge to assist them in the workplace)?				
16. Are there environmental risks associated with the AI System (including excessive pollution, or excessive energy or non-renewable resource consumption)?				
17. Contrary to Q16, does the AI System facilitate the environmentally and energy-efficient use of resources?				
18. Are there military or lethal uses for the proposed AI System? If so, answer Q19 and Q20 as appropriate. If no, go to Q21.				
19. In the case of military or lethal uses of the AI System: (a) Is the AI System fully autonomous? (b) is there a shutdown function triggered by designated personnel? (c) Is there effective human oversight in place?				
20. In the case of military and lethal uses of the AI System (a) is the AI System semi-autonomous? (b) Is there a shut down function triggered by designated personnel? (c) Is there effective human oversight in place?				
21. May the AI System be deemed to be a medical device or any other qualification that could entail application of other regulations (e.g. medical secrecy) that could modify its ethical perception?				

Risk Factors	Whether/How the Solution Addresses the Factors	Risk Rating	Mitigation Measures	Revised Risk Rating
22. Is there a risk that the AI System could violate ethical principles of beneficence and non-maleficence?				
23. Is there a risk that the AI System could have a negative impact on democratic and/or electoral processes?				
24. Is there a risk that the AI System could have a negative impact on judicial judgment and/or processes, legal procedural due process and/or access to justice?				
25. Is there a risk that the AI System could have a negative impact on learner pathways, assessment for attaining a qualification, assessment for a job or promotion, access to educational institutions and/or access to further learning opportunities?				
26. Is there a risk that the AI System could select, classify, or categorize or seek to ascertain a level of assurance concerning individuals (or groups of individuals) in such a manner as to deny them access to a good or service (or promote too high a barrier of entry resulting in effective exclusion) which is unreasonable and unjustifiable?				
		Average		**Average**

Principle 2: Accountability

Organisations that develop, make available or use AI Systems ought to be accountable for the consequences of their actions and shall designate an individual or individuals who are accountable for the organisation's compliance with the principles of this Policy Framework for Responsible AI or other adopted principles (including analogous principles that may be developed for a specific industry) with the objective of keeping humans behind the machines and AI Human centric.

Overview of Principle 2

The Organisation should ensure at all times that it remains accountable for the ethical and responsible deployment of AI Systems that the Organisation deploys, including by means of "human-in-the-loop" deployment.

Risk Factors	Whether/How the Solution Addresses the Factors	Risk Rating	Mitigation Measures	Revised Risk Rating
1. Is this AI System an expansion of a previous activity? If yes, determine whether a previous assessment has been done. If a previous assessment has been done, what has changed in this data activity and why (refer to previous assessment)?				
2. How experienced with tech projects is the team that will develop the AI System?				
3. What is the level of internal support, including financial, for the AI System?				
4. Who will be accountable within the organisation with regards to the AI System? Is there a central coordinating body? Who will be accountable within the organisation upon failure of the AI System, or upon production of adverse outcomes for its users?				
5. Will the staff be trained to use the AI System? Are the relevant personnel and/or departments fully aware of their roles and responsibilities? • This inquiry should account for the different types of staff and the different layers of personnel involved in the design of the AI System (e.g. management/oversight in addition to programming levels).				
6. What elements of the training and development "supply chain" have been outsourced? If handed off to a third party, are their services subject to the same levels of quality control as the Organisation?				
7. What are the roles played by the Organisation within the AI System pipeline (end-user, developer, data provider, etc.)?				
8. What will be the relation of the Organisation with end users once the AI System developed reaches the market (for instance, is AI System sold as a product or as a Software-as-a-Service)?				
9. To what extent does the AI System rely on third party data/systems input? How accountable are those third-party dependencies?				

Risk Factors	Whether/How the Solution Addresses the Factors	Risk Rating	Mitigation Measures	Revised Risk Rating
10. What is the maximum degree of autonomy that the AI System can reach?				
11. Identify all stakeholders that are affected by the AI System.				
12. How will the internal use of the AI System by the Organisation affect the roles and tasks of employees?				
13. Does the Organisation provide a method for individuals to access and correct personal information used in the AI System? How does this change if the data isn't deemed to be personal data (i.e. anonymized and not re-identifiable) but yet relates to a human?				
14. Does the AI System provide functionality allowing the user to "turn off" the app for a limited time?				
15. Does the AI System conform to industry or sector specific regulations given its deployment capabilities and its data source? (e.g. consumer protection, banking, health sector)				
16. Is there an independent commissioner committed to the review and control of such AI Systems? (e.g. governmental agency, designated official)				
17. Is a Privacy Policy available?				
18. Are the principles of necessity, proportionality and data minimization fully integrated?				
19. What privacy by design measures have been implemented?				
20. Are personal data that are being collected by the AI System used for any secondary purposes (after the "sunset" of the AI System)? Are secondary uses of data compatible with initial purposes, if any?				
21. How are transfers of data of the AI System outside of the EU/national/regional frontier organized?				
22. Have external QA/QC control methodologies been observed in the creation of the AI System (i.e. ISO 9001)?				
23. How will the AI model training and selection process be managed?				
24. Consider maintenance, monitoring, documentation and review of the AI models that have been deployed.				

Risk Factors	Whether/How the Solution Addresses the Factors	Risk Rating	Mitigation Measures	Revised Risk Rating
25. Consider the various degrees of human oversight in the decision-making process: a. **Human-in-the-Loop:** This model suggests that human oversight is active and involved, with the human retaining full control and the AI only providing recommendations or input. Decisions cannot be exercised without affirmative actions by the human, such as a human command to proceed with a given decision. (NB: Considering here also the concept of "Human in the Loophole" where there is automation bias, anchoring or confirmation bias in respect of the human operative. The human essentially affirming the AI outcome without critically assessing whether it is correct or not.) b. **Human-out-of-the-Loop:** This model suggests that there is no human oversight over the execution of decisions. AI has full control without the option of human override. c. **Human-over-the-Loop:** This model allows humans to adjust parameters during the execution of the algorithm.[1]				
26. What are the rights and interests at stake when the AI System makes an automated decision?				
		Average		**Average**

Principle 3: Transparency and Explainability

Organisations that develop, make available or use AI Systems, and any national laws or industry standards that govern such use, shall ensure that such use is transparent and that the decision outcomes of the AI System are explainable.

Overview of Principle 3

- Organisations that make available or use an AI System in decision-making processes must disclose certain meaningful information to enable individuals the opportunity to choose whether to proceed, and, if so, to understand the decision and decide whether to contest it.
- The intensity of the transparency and explainability disclosure obligations will depend upon a variety of factors, including the nature of the data involved, lack of human participation in the decision-making, result of the decision and its consequences for the affected individual.
- Organisations that develop AI Systems should ensure that the system architecture, algorithmic logic, data sets, testing methods, and all related development and operational policies and procedures serve to embed transparency and explainability by design.

Risk Factors	Whether/How the Solution Addresses the Factors	Risk Rating	Mitigation Measures	Revised Risk Rating
1. Has a governance methodology been implemented to apply transparency and explainability by design principles throughout the development lifecycle?				
2. Have developmental and operational policies procedures and controls been implemented pursuant to such methodology?				
3. Have internal controls been developed pursuant to such policies and procedures?				
4. How did the selection of the system architecture and algorithmic model take transparency and explainability into account?				
5. How did the selection of data sets to train and test the AI System take transparency and explainability into account?				
6. Do terms and conditions apply to those individuals who may wish to access and use the AI System ("Terms of Use")?				
7. Are the Terms of Use clearly and prominently displayed?				
8. Are there any limitations on accessing the Terms of Use (e.g. a registration process)?				
9. What steps were taken to ensure the Terms of Use are accurate?				
10. What steps were taken to ensure the Terms of Use are objectively clear and readily understandable to a layperson?				
11. Do the Terms of Use vary based upon the level of sophistication or other attributes of a user? If so, how?				

Risk Factors	Whether/How the Solution Addresses the Factors	Risk Rating	Mitigation Measures	Revised Risk Rating
12. Do the Terms of Use apply a layered disclosure approach to allow interested individuals the ability to obtain more information about the AI System?				
13. Do the Terms of Use provide meaningful information regarding the fact that an AI System is being used in a decision-making process? If so, how?				
14. Do the Terms of Use provide meaningful information regarding the intended purpose(s) of the AI System? If so, how?				
15. Do the Terms of Use provide meaningful information regarding the types of data sets that are used and generated by the AI System? If so, how?				
16. Do the Terms of Use provide meaningful information regarding whether and to what extent the decision-making process may include human participation? If so, how?				
17. Are prior versions of the Terms of Use publicly available?				
18. Is there a process to periodically review and update the Terms of Use?				
19. Is there a process to periodically assess whether users understand the Terms of Use?				
20. How are the results of the AI System made available to users?				
21. When are the results of the AI System made available to users?				
22. At such time, what information is provided regarding the algorithmic logic of the AI System?				
23. At such time, what information is provided to understand the decision/outcome?				
24. At such time, what information is provided regarding how to contest the decision/outcome?				
25. At such time, what information is provided regarding the traceability or auditability of the AI System?				
26. At such time, what information is provided regarding the testing methods of the AI System?				
27. Are any other disclosures made with respect to the transparency and explainability of the AI System (e.g. videos, icons, symbols, white papers, dashboards, or counterfactual interfaces)?				

Risk Factors	Whether/How the Solution Addresses the Factors	Risk Rating	Mitigation Measures	Revised Risk Rating
28. Does the disclosure of any information listed in this section change depending on the nature of the data involved (e.g. if sensitive personal data is used by the AI System)? If so, how?				
29. Does such disclosure change depending on the lack of human participation in the decision-making? If so, how?				
30. Does such disclosure change depending on the result of the decision and its consequences for the user (e.g. if legal or human rights are materially affected)? If so, how?				
31. Is the AI System periodically audited or assessed with respect to transparency and explainability, either internally or by an independent third party?				
		Average		**Average**

Principle 4: Fairness and Non-Discrimination

Organisations that develop, make available or use AI Systems and any national laws that regulate such use shall ensure the non-discrimination of AI outcomes, and shall promote appropriate and effective measures to safeguard fairness in AI use.

Overview of Principle 4

- The use of the AI System should be non-discriminatory in terms of accessibility. The AI System should be accessible also to people with disabilities (such as, for instance, limited visual capacity).
- Decisions based on the AI System should be fair and non-discriminatory, judged against the same standards as decision-making processes conducted entirely by humans, and where possible seek to achieve a higher standard of fairness and non-discrimination.
- AI development should be designed to prioritize fairness. This would involve addressing algorithms and data bias from an early stage with a view to ensuring fairness and non-discrimination throughout the whole AI System lifecycle.

Risk Factors	Whether/How the Solution Addresses the Factors	Risk Rating	Mitigation Measures	Revised Risk Rating
1. Is the use of the AI System voluntary, incentive-based or compulsory?				
2. Is the AI System following a deterministic approach as opposed to a probabilistic model?				
3. Is the AI System making automated decisions affecting the rights and interests of individuals or businesses? • It should notably be considered whether the AI System may have consequence for the user to suffer differential treatment which would otherwise be prohibited under any applicable law. • Also consider whether the AI may hamper the effective enforcement of existing laws meant to protect fundamental rights, due to unperceived bias (e.g. candidates of a certain sex, disability or ethnicity may not see certain job vacancies in the first place) or bias which is difficult to challenge without appropriate documentation about how the system works; or about the goals it pursues (e.g. automatic denial or recovery of social security benefits).[2]				
4. Does the Organisation understand the lineage of data (where the data originally came from, how it was collected, curated and moved within its Business Unit/Division, and how its accuracy is maintained over time)? Consider keeping a data provenance record.				

Risk Factors	Whether/How the Solution Addresses the Factors	Risk Rating	Mitigation Measures	Revised Risk Rating
5. Is the data high quality data? The following factors should be assessed: • the accuracy of the dataset, in terms of how well the values in the dataset match the true characteristics of the entities described by the dataset; • the completeness of the dataset, both in terms of attributes and items; • the veracity of the dataset, which refers to how credible the data is, including whether the data originated from a reliable source; • how recently the dataset was compiled or updated; • the relevance of the dataset and the context for data collection, as it may affect the interpretation of and reliance on the data for the intended purpose; • the integrity of the dataset that has been joined from multiple datasets, which refers to how well extraction and transformation have been performed; • the usability of the dataset, including how well the dataset is structured in a machine-understandable form; • the usability of any personal information contained within the data sets, including with regards to obtaining any requisite consents; and • human interventions, e.g. if any human has filtered, applied labels, or edited the data.				
6. Is the data used for the training of the AI System representative of the population about which the AI System will make decisions (data accuracy, data quality and data completeness)?				
7. Does the Organisation have an established and robust selection process in relation to the datasets training the AI System? For example, are there minimum requirements as to the diversity and quality of the datasets used?				
8. Does the AI System use different datasets for training, testing and validation?				

Risk Factors	Whether/How the Solution Addresses the Factors	Risk Rating	Mitigation Measures	Revised Risk Rating
9. Consider minimizing inherent bias: • **Selection Bias:** This bias occurs when the data used to produce the AI System are not fully representative of the actual data or environment that the AI System may receive or function in. Common examples of selection bias in datasets are omission bias and stereotype bias. • **Measurement Bias:** This bias occurs when the data collection device causes the data to be systematically skewed in a particular direction. • **Weighting Bias:** This bias occurs when the data used by the AI Solution are attributed differing weights in producing the relevant outcome. The datasets might be afforded greater or lesser value, which might be arbitrarily or inaccurately awarded. • **The following factors should be assessed** (amongst others): – the frequency with which the dataset is reviewed and updated; – representativeness of the dataset to the end-user demographic and desired outcomes; – the diversity of the dataset, and the variety of sources from which the data has been collected (i.e. numeric, text, audio, visual, transactional, etc.); and – the usability of different datasets, including how those datasets have been matched and cleaned so that relational datasets can be correlated and linked.				
10. Is there rigorous testing of the AI System, both before use and periodically afterwards, to ensure that there is no disparate impact on a protected class of individuals?				
11. How are "edge cases" managed by the AI System?				
12. Does the Organisation have in place a system to respond to and resolve situations in which the AI System produces discriminatory or unfair outcomes? • This should encompass the Organisations' capacity to assess and identify biased datasets, potential relief measures provided to end users and any scope to redesign the AI System.				
13. What methodologies have been applied and used in the training of the AI System?				
14. Does the AI System have a fixed learning phase followed by a static use phase or does it continuously improve? If the latter, how are improvements filtered for bias, quality, etc.?				

Risk Factors	Whether/How the Solution Addresses the Factors	Risk Rating	Mitigation Measures	Revised Risk Rating
15. What are the risks of bias existing or occurring in 1) the algorithm, 2) the training data, 3) the human designers and developers, and 4) end users?				
16. What are the reputational risks for the Organisations of the AI System making biased automated decisions?				
		Average		**Average**

Principle 5: Safety and Reliability

Organisations that develop, make available or use AI Systems shall adopt design regimes and standards ensuring high safety and reliability of AI Systems on one hand while limiting the exposure of developers and deployers on the other hand.

Overview of Principle 5

- Organisations developing, making available or using AI Systems define the relevant set of ethical and moral principles underpinning the AI System to be developed, deployed or used taking into account all relevant circumstances.
- Organisations should test AI Systems thoroughly to ensure that they reliably and robustly adhere, in operation, to the underpinning ethical and moral principles and have been trained with data which are curated and are as "error-free" and "bias-free" as practicable, given the circumstances.

Risk Factors	Whether/How the Solution Addresses the Factors	Risk Rating	Mitigation Measures	Revised Risk Rating
(Pre-) Design, Development & Testing				
1. Is there a clearly defined set of relevant ethical and moral principles in place on the basis of which the AI System is intended to operate, such taking into account all relevant circumstances? a. Have all local standards been identified and taken into account e.g. in relation to geographical, religious and/or social considerations and traditions? b. Are the underpinning ethical and moral principles periodically validated to ensure on-going accurateness, starting with a validation prior to the design and development of the AI System?				
2. Have ethical and moral appropriateness considerations been translated into (technical and/or functional) boundaries affecting the outcome of the AI System's use (e.g. its decision-making powers)? What is the impact of this on the general accuracy of the outcome of the AI System's use?				

Risk Factors	Whether/How the Solution Addresses the Factors	Risk Rating	Mitigation Measures	Revised Risk Rating
3. Have safety and reliability risk scenarios been identified, both for the AI System's users and beyond (e.g. potentially indirectly affected stakeholders or society at large), including associated risk metrics and risk levels, in relation to: a. the quality and performance of the AI System itself (e.g. design faults, technical defects, low level of accuracy, unintended self-learning capabilities); b. the data and assumptions used to develop and train the AI System (e.g. preventing data that are not up-to-date, incomplete and/or non-representative); c. any possible (harmful) use of the AI System or the outcome thereof (e.g. over-reliance, human attachment, addictive user behaviour and manipulation of user behaviour), including any malicious, inappropriate or unintended (dual) use; and d. the safety and reliability expectations of the users and their level of sophistication.				
4. Has a definition been set of what is considered to be a safe and reliable AI System, and is this definition commonly used and implemented throughout the full lifecycle of design, development, deployment, operation and use of the AI System? a. Have quantitative analysis or metrics been applied to measure and test the applied definition? b. Are there regulatory requirements that impact the above definition of safety and reliability (e.g. medical devices regulations)?				
5. Have clear fault tolerance requirements been set that are considered acceptable in relation to the intended outcome of the AI System's use? If yes, what is the basis for setting these fault tolerance requirements (e.g. a legacy solution that the AI System will be replacing)?				
6. Has the AI System been assessed to determine whether (and if so, the extent to which) it is also safe for, and can be reliably used by, those with special needs or disabilities or those at risk of exclusion?				
7. Are all safety and reliability considerations as addressed in aforementioned questions expressed in the design and development documentation in sufficient detail?				
8. How is the AI System's testability and auditability facilitated?				

Risk Factors	Whether/How the Solution Addresses the Factors	Risk Rating	Mitigation Measures	Revised Risk Rating
9. Is the testing procedure aligned to the appropriate levels of safety and reliability as needed, taking into account the safety and reliability considerations expressed in the design and development documentation? Does the testing procedure also accommodate for testing of the AI System in "edge cases" (use scenarios that are unlikely to occur but are nonetheless possible)?				
Deployment and Operation				
10. Has a "pilot" deployment been considered to enable testing and refining the operation of the AI System and to expedite the completion of the AI System improve its safety and reliability? If yes, has this pilot been limited in time and users, have users been informed about the specifics of the pilot, and is it possible to safely abort upon short notice?				
11. Are there any specific human oversight and control measures in place that reflect the safety and reliability risks of the AI System, given the degree of self-learning and autonomous features of the AI System?				
12. What procedures are in place to ensure the explainability of the AI System's decision-making process during operation?				
13. How is the ongoing auditing of the AI System's safety and reliability organised and facilitated, internally as well as by independent third parties? Aside from exception reporting, does this also include failure analysis to determine causes or fixes for any problems? Is safety audited separately from reliability?				
Users				
14. Are users informed on: a. the (technical and/or functional) boundaries implemented to affect the outcome of the AI System's use; b. the potential safety and reliability risks of the AI System to the users (e.g. the level of accuracy of the AI System to be expected by users); and c. the duration of coverage and schedules timeframes for security and other updates to improve the safety and/or reliability of the AI System?				
15. Are appropriate training materials on how to ensure a safe and reliable use of the AI System provided to users within the limitations communicated for such safe and reliable use?				

Risk Factors	Whether/How the Solution Addresses the Factors	Risk Rating	Mitigation Measures	Revised Risk Rating
Resilience				
16. Is the AI System certified for cybersecurity in accordance with an international certification scheme, or is it otherwise demonstrably compliant with specific security standards?				
17. To what extent is the AI System exposed to and protected against potential cyber-attacks over its lifecycle? What potential forms of attacks, types of vulnerabilities and potential entry points for attacks have been taken into account in this respect?				
18. Is the AI System subjected to routing penetration testing and/or red team testing?				
Risk Monitoring, Alteration & Control				
19. Is there a process in place to continuously measure and assess safety and reliability risks in accordance with the risk metrics and risk levels defined in advance for each specific use case?				
20. Are there procedures and/or measures in place that ensure comprehensive and transparent investigation of adverse, unanticipated and/or undesirable alterations to or outcomes of the AI System, in particular in the event of resulting harm to the safety of its users or beyond (e.g. to society at large), and that mitigate any risks of such resulting harm occurring?				
21. Is there a mechanism in place that allows for designers, developers, users, stakeholders and third parties to (anonymously) flag/report vulnerabilities and other issues related to the safety and reliability of the AI System?				
22. Are there tested failsafe fall-back plans to address the AI System's errors of whatever origin, including governance procedures to trigger them?				
23. Is the AI System designed in such a way (e.g. by including a 'stop button') that it can safely and elegantly abort the deployment and/or operation of the AI System when needed without catastrophic results for the users and beyond?				
24. How are the results of all risk assessment, risk management and risk control procedures in relation to safety and reliability of the AI System factored into necessary or desirable alterations of (the design of) the AI System? How is this process documented?				
		Average		**Average**

Principle 6: Open Data, Fair Competition and Intellectual Property

Organisations that develop, make available or use AI Systems and any national laws that regulate such use shall, without prejudice to normal rules of intellectual property and privacy:

(a) foster open access to, and the portability of, datasets (where privately held), especially where such datasets are deemed significant and important or advance the 'state of the art' in the development of AI Systems;

(b) ensure that data held by public sector bodies are, in so far as is reasonably practicable, portable, accessible and open; and

(c) encourage open source frameworks and software for AI Systems which could similarly be regarded as significant and important and advance the 'state of the art.'

AI Systems must be developed and made available on a "compliance by design" basis in relation to competition/antitrust law.

Principle 8: Intellectual Property

Organisations that develop, make available or use AI Systems should seek to strike a fair balance between benefiting from adequate protection for the intellectual property rights for both the AI System and the AI output and allowing availability for the wider societal benefit. Governments should investigate how AI Systems and AI-created output may be afforded adequate protection whilst also ensuring that the innovation is sufficiently disclosed to promote progress.

Overview of Principles 6 and 8

- The Organisation should assess how its AI System and its outputs can be used in other situations, contexts or other applications (which differ from the original use case or original design goal) or by other Organisations.
- Private organisations should foster open access and portability of datasets.
- Public sector bodies must ensure that data held by them are portable, accessible and open if reasonably practicable.
- Organisations should encourage open source frameworks and software to advance the 'state of the art' for AI solutions.
- The Organisation should take into account competition law when developing the AI System.
- Organisations must be allowed to protect rights in their AI Systems. However, care needs to be taken not to take steps which will amount to overprotection, as this could prove detrimental to "state of the art" development.

Risk Factors	Whether/How the Solution Addresses the Factors	Risk Rating	Mitigation Measures	Revised Risk Rating
1. Does the Organisation offer easy portability of its privately held datasets? If so, is it clear for what purposes the datasets may be transferred and whether there will be any remuneration for transfer?				
2. Does the Organisation foster open access to its privately held datasets? If so, is it clear who can access the datasets, for what purposes the datasets can be used and whether there will be any remuneration for granting access?				
3. If datasets are made available by a public sector body, how is it ensured that the data is portable, accessible and open? And if so, is it clear who can do what with the data?				

Risk Factors	Whether/How the Solution Addresses the Factors	Risk Rating	Mitigation Measures	Revised Risk Rating
4. What categories of data are required for the use of the AI System? Have all rights to data from third parties been cleared and agreed in a license agreement?				
5. What categories of data will be produced by the AI System? Will the resulting data be made available to third parties? If yes, what type of licensing arrangement is appropriate for providing data resulting from the AI System to third parties to ensure a fair balance between the Organisation's commercial use of the data and promoting open access to data?				
6. What is the scope of interoperability with other tech solutions offered by the same or other providers?				
7. Is the data generated by the AI System reusable in the public interest (data for good projects)?				
8. What are the ownership or intellectual property rights attaching to the AI System?				
9. Are there any compulsory licensing or patent rights issues relating to the AI System?				
10. Have the intellectual property rights attaching to the AI System been made publicly available (i.e. turning the underlying code into an open source program)?				
11. Is competition law compliance taken into account when developing the AI System, such as designing to reduce the risk of the AI System using anti-competitive behaviour to reach its purpose ("compliance by design")?				
12. Are there any other project-specific risks relating to Principles #6 and #8, which need to be taken into account in the RAIIA?				
		Average		**Average**

Principle 7: Privacy

Organisations that develop, make available or use AI Systems and any national laws that regulate such use shall endeavour to ensure that such AI Systems are compliant with privacy norms and regulations, taking into account the unique characteristics of AI Systems and the evolution of standards on privacy.

Overview of Principle 7

The organisation should consider implementing operational safeguards to protect privacy such as privacy by design principles that are specifically tailored to the specific features of the deployed AI System.

Risk Factors	Whether/How the Solution Addresses the Factors	Risk Rating	Mitigation Measures	Revised Risk Rating
1. Consider if the data is provided by the individual (originated in direct action taken by the individual) and whether: • The data is initiated (the product of individuals taking an action that begins a relationship) • The data is transactional (created when the individual is involved in a transaction) • The data is posted (created when individuals proactively express themselves)[3]				
2. Consider if the data is observed (created as the result of individuals being observed and recorded), whether: • The data is engaged (instances in which individuals are aware of observation at some point in time) • The data is not anticipated (instances in which individuals are aware there are sensors but have little awareness that sensors are creating data pertaining to the individuals) • The data is passive (instances in which it is very difficult for the individuals to be aware they are being observed and data pertaining to observation of them is being created)				
3. Consider if the data is derived (created in a mechanical fashion from other data and becomes a new data element related to the individual), whether: • The data is computational (creation of a new data element through an arithmetic process executed on existing numeric elements) • The data is notational (creation of a new data element by classifying individuals as being part of a group based on common attributes shown by members of the group)				
4. Consider if the data is inferred (product of a probability-based analytic process), whether: • The data is statistical (the product of characterization based on a statistical process) • The data is advanced analytical (the product of an advanced analytical process)				

Risk Factors	Whether/How the Solution Addresses the Factors	Risk Rating	Mitigation Measures	Revised Risk Rating
5. How was the data used by the AI System collected and stored? Was the data transferred by third parties or will the data be transferred to third parties? • Consider whether preprocessing activity has been done on the data before the analysis and whether it would have affected the accuracy and appropriateness of individuals.				
6. Who were the data subjects? What type of information was collected about them? What is the scope of the consents obtained?				
7. Is sensitive data collected? If so, are there higher standards being adopted for protection of this kind of data?				
8. Beyond the data subjects' privacy, may the privacy of an identified group be at risk?[4]				
9. Are there viable alternatives to the use of personal information (e.g. anonymization or synthetic data)? If so, what mechanisms/techniques are implemented to prevent from re-identification?				
10. Are there procedures for reviewing data retention and performing destruction of data used by the AI System? Are there oversight mechanisms in place?				
11. What is the nature of the Organisation's relationship with the data subjects? How much control will they have? Would they expect you to use their data in this way?[5]				
12. Do they include children or other vulnerable groups? Are there prior concerns over this type of processing or security flaws?				
13. What is the Organisation's lawful basis for processing personal information? What measures does the Organisation take to ensure compliance?				
		Average		**Average**

4. Risk Assessment Summary

This section describes the risks you've identified through the RAIIA process and how you propose to mitigate and manage those risks. It can be useful to link this back to the principles to show why these risks and the proposed actions are relevant. Document the risks in line with any existing risk management processes the Organisation has—it will be more efficient than trying to run a separate process.

5. Risk Mitigation Action Plan

This section describes how you propose to mitigate and manage the risks previously described. In some cases, it may be helpful to categorize these actions into areas such as:

- Governance
- People
- Process
- Technology

Please provide details of all such strategies. Also, please identify the likelihood (low, medium, or high) of this risk happening and the degree of impact it would have on individuals if it occurred. You can use the form of the table below.

	Risk	Mitigation Strategy	Likelihood	Impact
1.				
2.				
3.				
4.				
5.				

Endnotes

1 Singapore's Model Artificial Intelligence Governance Framework, Second Edition, https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf.

2 Comp. Report of the United Nations Special rapporteur on extreme poverty and human rights, published 11 October 2019, https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25156.

3 Hong Kong Privacy Commissioner and the Information Accountability Foundation, 2018.

4 In addition to assess who were the data subjects (i.e. question 2), it appears relevant to integrate the "group privacy" concept herein. It corresponds to the view that the protection of the privacy of a group should also be a goal of privacy regulation, in response to advances in big data technology. So far, privacy regulation are mainly centered on identifiable individuals, but theoreticians of the "group privacy" concept state that there are also risks for privacy, resulting from the assumption that if the privacy of individuals is taken care of, the privacy of groups will take care of itself. This warrants the philosophical exploration of theories of group privacy, which conceptualize group privacy as the privacy of a group which is not achieved, automatically, by protecting the individual privacy of all members of a group. As an example, the impact on medical staff might be assessed as well. They are indeed not necessarily data subjects, but they may be indirect stakeholders of the AI Solution, and regarding their role to play in the collection/processing of data, the question arises whether their privacy is not also at risk. See List of Sources for RAIIA Template, following, for resources on the subject.

5 "Sample DPIA template," online: Information Commissioner's Office, https://gdpr.eu/wp-content/uploads/2019/03/dpia-template-v1.pdf.

List of Sources for RAIIA Template

1. Council of Europe guidelines to assess algorithms and automation to prevent against human rights breaches, online: https://search.coe.int/cm/pages/result_details.aspx?objectid=09000016809e1154.
2. Data protection impact assessments, online: Information Commissioner's Office, https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/.
3. Data Protection Impact Assessments: Data Protection Commission, online: Data Protection Commission, https://www.dataprotection.ie/en/organisations/know-your-obligations/data-protection-impact-assessments.
4. Ethical Accountability Framework and Data Stewardship Accountability, Data Impact Assessments and Oversight Models, a joint publication of the Hong Kong Privacy Commissioner and the Information Accountability Foundation (2018), online: https://www.pcpd.org.hk/english/resources_centre/publications/surveys/surveys.html.
5. EU HLEG Ethics Guidance on Trustworthy AI, online: https://ec.europa.eu/futurium/en/ai-alliance-consultation.
6. EU HLEG Report on Liability for AI and other digital emerging technologies, online: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=63199.
7. High-Level Expert Group on Artificial Intelligence: Assessment List for Trustworthy Artificial Intelligence (ALTAI).
8. EU Commission, White Paper on Artificial Intelligence: a European approach to excellence and trust (19 February 2020): https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

9. EU Commission, Proposal for a legal act of the European Parliament and the Council laying down requirements for Artificial Intelligence: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Requirements-for-Artificial-Intelligence.
10. Report of the United Nations Special rapporteur on extreme poverty and human rights, (11 October 2019): https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25156.
11. G20 Ministerial Statement on Trade and Digital Economy, (9 June 2019), online: Munk School of Global Affairs and Public Policy, http://www.g20.utoronto.ca/2019/2019-g20-trade.html.
12. Individual vs Group Privacy, (20 March 2019), online: http://www.ithappens.nu/individual-vs-group-privacy/.
13. ITechLaw Responsible AI framework, https://www.itechlaw.org/ResponsibleAI.
14. Modernised Convention for the Protection of Individuals with Regard to the Processing of Personal Data, https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=09000016807c65bf.
15. OECD principles, online: https://www.mofa.go.jp/files/000486596.pdf.
16. Opinion 05/2014 on Anonymisation Techniques, (10 April 2014), online: European Data Protection Board (former WP29), https://www.pdpjournals.com/docs/88197.pdf.
17. Sample DPIA template, online: Information Commissioner's Office, https://gdpr.eu/wp-content/uploads/2019/03/dpia-template-v1.pdf.
18. Singapore's Model Artificial Intelligence Governance Framework, Second Edition, https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf.
19. Smith, Andrew. "Using Artificial Intelligence and Algorithms", (8 April 2020), online: Federal Trade Commission, https://www.ftc.gov/news-events/blogs/business-blog/2020/04/using-artificial-intelligence-algorithms.


ANTHONY WONG


astrea


BAS
BELÉN ARRIBAS
ABOGADA


Beyond Reach Consulting Limited
Thinking Beyond Reach to Reach Beyond. Applying Digital Ethics by Design into your DNA


bpv HÜGEL


CMS
Law.Tax


DAC BEACHCROFT


delacruz beranek
RECHTSANWÄLTE • ATTORNEYS AT LAW


DW
Deeth Williams Wall


ERVIN COHEN & JESSUP LLP


FENWICK


FRORIEP


GATTAI, MINOLI, AGOSTINELLI,
PARTNERS
STUDIO LEGALE


HOUTHOFF


KVALE


Lawways
Avocats


MARVAL
O'FARRELL
MAIRAL


MKW
MAURIEL KAPOUYTIAN WOODS LLP


mccarthy tetrault


Osborne Clarke


Seyfarth


Tanner De Witt
solicitors


TaylorWessing


TIMES
ATTORNEYS


TRILEGAL


UGHI E NUNZIANTE
STUDIO LEGALE


VANDOORNE


Weitnauer


WP
DR. WIDMER & PARTNER · RECHTSANWÄLTE
DR. WIDMER & PARTNERS · ATTORNEYS-AT-LAW


WP THOMPSON
INTELLECTUAL PROPERTY

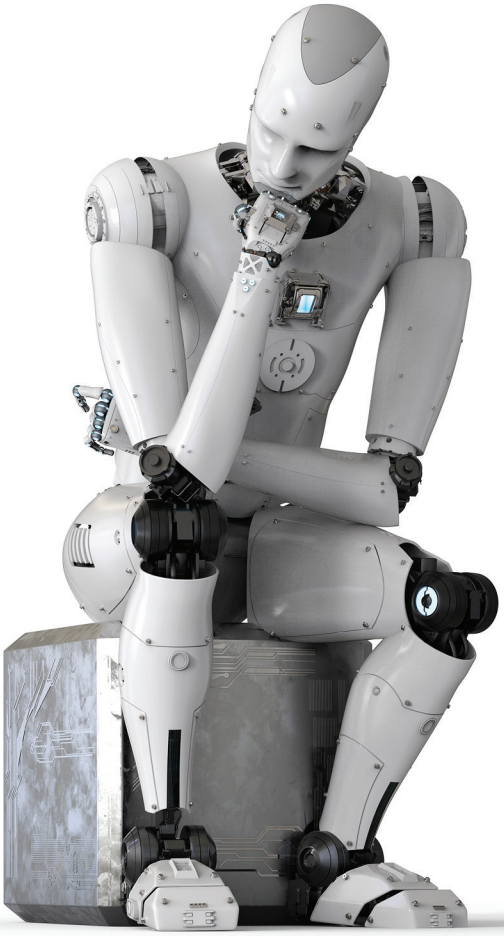
Responsible AI

A GLOBAL POLICY FRAMEWORK

ITechLaw is the leading global organisation for legal professionals focused on technology and law. It has been serving the technology law community worldwide since 1971.

ITechLaw has a global membership base representing more than 70 countries on six continents. Its members reflect a broad spectrum of expertise in the technology law field. Our mission is to create unparalleled opportunities for international networking and for the exchange of knowledge with experts and colleagues around the world.

ITechLaw is a thought leader in many areas of technology law, including as regards the responsible development, deployment and use of artificial intelligence.



ITECHLAW®
INTERNATIONAL TECHNOLOGY LAW ASSOCIATION

itechlaw.org

ISBN 978-1-7339931-1-1