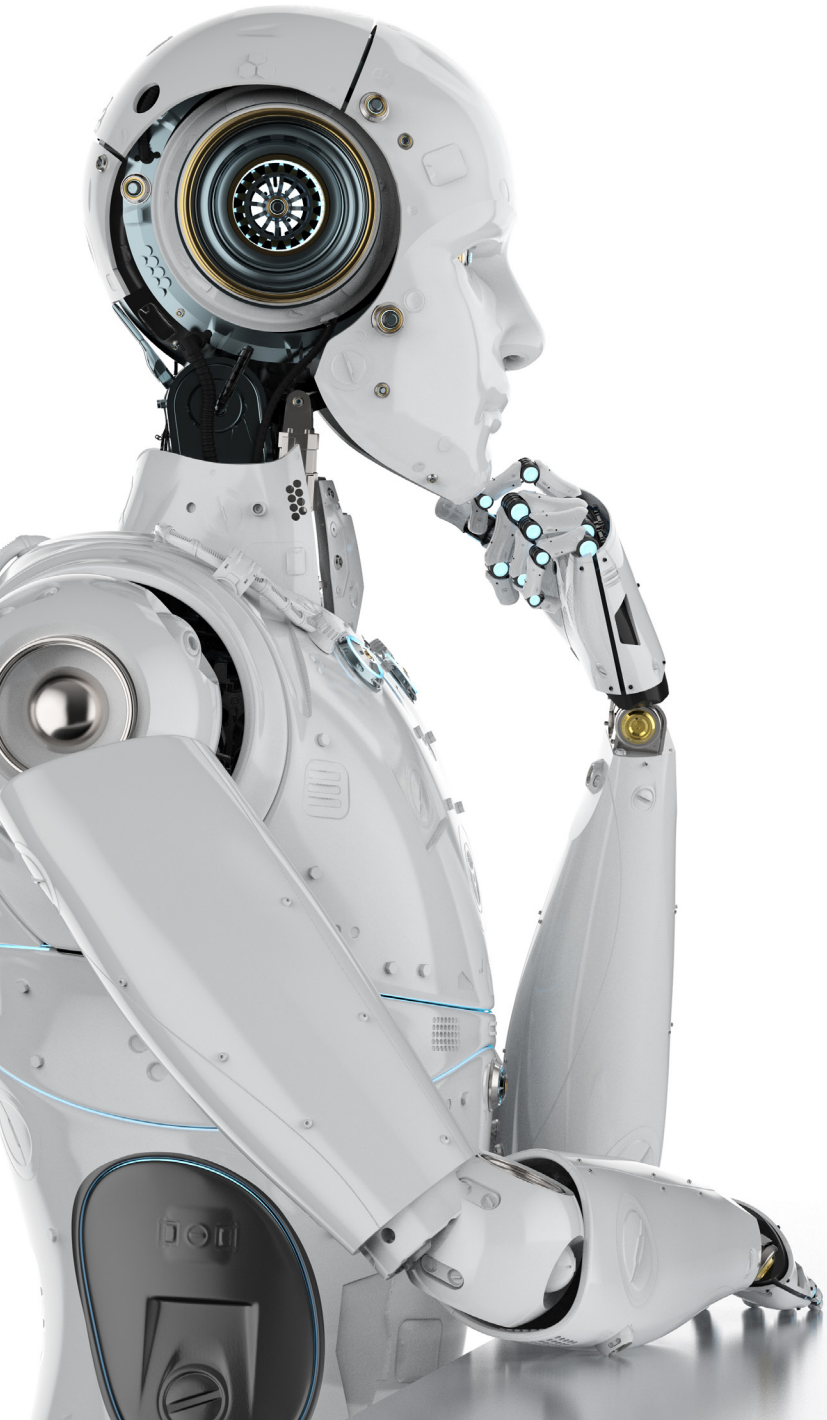
Responsible AI

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://optictechnology.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/.