Chapter 1
AI: the value of precaution and the need for human control

Aída Ponce Del Castillo

'*Whose duty is it in today's complex societies to foresee or forestall the negative impacts of technology, and do we possess the necessary tools and instruments for forecasting and preventing harm?*'

Sheila Jasanoff, Pforzheimer Professor of Science and Technology Studies, Harvard Kennedy School

'*The aim is quite simple. Let's use it more. Let's have artificial intelligence everywhere where it makes a difference.*'

European Commission Executive Vice-President Margrethe Vestager, at the 2021 Data Science & Law Forum

1. Introduction

When new technologies emerge, two opposing governance approaches may arise. One favours stringent regulation to safeguard society from unanticipated hazards while the other prioritises the promotion of technology deployment by eschewing what is often seen as expensive and innovation-stifling regulation (Kaal 2016; Cortez 2019; Mandel 2020). Between these two extremes, various governance approaches can be developed. These will include governance through international human rights, or through hard law (which includes risk-based regulation) or soft law.

Whatever approach is chosen, the timing of governance interventions is crucial. If interventions and course corrections are made early, they are likely to be less expensive and easier to carry out. However, the full implications of emerging technology and the need for change might not yet be fully understood. Delaying intervention until it becomes necessary can result in more challenging, time-consuming and costly course corrections (Collingridge 1982).

Artificial intelligence (AI), similarly to other high-risk emerging technologies such as biotechnology, blockchain, synthetic biology, metaverse environments and nanotechnology, presents certain key attributes: radical novelty; relatively fast growth; coherence over time; a prominent impact on society and the economy; and uncertainty and ambiguity (Rotolo et al. 2015). The European Commission (EC), in its proposed regulation on machinery products, mentions additional attributes including data dependency, opacity, autonomy and connectivity, recognising that these can increase both the probability of harm and its impact as well as negatively affect the safety of any machinery that integrates AI (European Commission 2021).

Compared to other high-risk emerging technologies, AI also raises several unique concerns related to possible bias and discrimination, the preservation of democratic values, the need to render automated decision-making more explicit (Taeihagh et al. 2021), widening inequalities, the impact of bad data, the protection of data privacy and the prevention of mass surveillance (Zuboff 2019; Zhang et al. 2021).

Against this high-risk background of emerging, rapidly evolving, uncertain and high-impact technology, the option of shifting the point of initial governance to an earlier stage of technological development is both necessary and valid. The risk control approach presently employed for established and clearly defined technologies, employing risk assessment followed by risk management to maintain exposure levels below the acceptable, is inadequate for emerging high-risk technologies: the quantitative data required is limited or unavailable; and the potential consequences of using the technology cannot be comprehensively listed (Linkov et al. 2018). Establishing a proactive, collaborative and flexible form of precautionary governance that evolves in conjunction with the technology may improve our prospects of safeguarding society from AI's potential impacts (Mandel 2013, 2020).

Such an approach, which establishes an initial point of governance at an early stage, ought to incorporate two essential principles: firstly, the legal principle of precaution, which has demonstrated its effectiveness in managing the risks associated with emerging and unpredictable technologies; and secondly, the principle of human-in-control.

This is crucial if we acknowledge that AI is an extension of conventional automation and operates under the same general principle – 'If it is technically and economically feasible to automate a function, automate it' (Billings 1996: 9) – which triggers an obvious question: can automation exist without human control?

2. AI case studies: why human intervention is necessary

The risks posed by AI are not theoretical and academic researchers have identified many. Sometimes, however, reality speaks louder than words. The four case studies described below illustrate how AI and non-AI algorithms have delivered wrong outputs or 'recommendations' and had a significant impact on the lives of thousands of vulnerable individuals: people have lost their right to welfare benefits, families have been torn apart, students have been assigned wrong grades, etc.

Case 1 UK A-level grades algorithm: the teachers feeding the system

In 2020, the UK Education Ministry made a decision to cancel exams due to Covid-19 and sought alternative solutions to assign grades to students. The two solutions identified were: (a) rely entirely on teacher assessed grades; or (b) standardise the results (Kippin and Cairney 2022).

Tool: the Office of Qualifications and Examinations Regulation (Ofqual) chose option 2 and designed and implemented an algorithm to grade students (Office for Artificial Intelligence 2020). Teachers were asked to supply for each student and for every subject an estimated grade and a ranking, compared with every other student at the school within the same estimated grade. A-Level students were awarded grades generated by the algorithm which 'looked at the historical grade distribution of a school and then decided a students' grade on the basis of their ranking' (Kolkman 2020).

Consequences: tens of thousands of school pupils received grades lower than they had anticipated (Kolkman 2020). Ofqual's grading algorithm also affected many schoolteachers. Government officials argued that 'basing grades on teacher estimates alone would damage the credibility of this year's results compared to previous years' and the whole issue turned into a public scandal (Coughlan 2020).

Role of the human: for policy reasons, the algorithm was preferred to the teacher. As the report from Ofqual seems to suggest, teachers were interviewed about the grading process and requested to feed the algorithm. However, it is unclear if the teachers who were interviewed were informed about the intention to use an algorithm for the A-level grading process (Office for Artificial Intelligence 2020).

Case 2 The Dutch childcare benefits scandal (*toeslagenaffaire*): where were the supervisors?

In 2003, the Dutch authorities developed an automated welfare fraud detection system called Systeem Risico Indicatie (SyRI) (van Bekkum and Borgesius 2021). Used by the Dutch Tax and Customs Administration, the system used algorithms in which 'foreign sounding names' and 'dual nationality' were used as indicators of potential fraud (European Parliament 2022). The existence of this system was reported by the media in 2020 and an investigation by the Dutch Data Protection Authority found that the algorithms were discriminatory, among other things because they took into account variables such as someone having a second nationality.

Tool: a risk detection algorithm to process the social security documents of individuals applying for childcare benefits, alongside another machine learning tool, namely a risk-scoring algorithm to automate the selection of childcare allowance recipients. The system derived risk factors based on the analysis of historical data in order automatically to process the documentation and select welfare recipients for audits. Then, officials scrutinised those claims with the highest risk label (Hadwick and Lan 2021).

Consequences: the output of the algorithms became biased and unfair as the algorithm concluded that non-Dutch welfare recipients were more prone to fraud. Some 26,000 parents were mistakenly accused by the Dutch tax authorities of fraudulently claiming child allowance over several years from 2012, while 10,000 families were forced to repay tens of thousands of euros, in some cases leading to unemployment, bankruptcies and divorces (Henley 2021).

Role of the human: the decision to cut a family off from benefits payments should have gone through an extensive review process. The choices were left to algorithms. The key question here is the role of supervisors as 'supervisors are still responsible for their work, even if part of it is performed by a computer' (Ten Seldam and Brenninkmeijer 2021).

Case 3 The French *Foncier Innovant* system to identify swimming pool fraud: surveyors not consulted

In 2021, the State Public Finance Department developed an AI tool to detect undeclared outbuildings and swimming pools as part of a plan to update and measure all buildings on the national territory. The State's objective was to update available maps, improve their quality and facilitate the collection of taxes to be paid by homeowners (Direction générale des Finances publiques 2022).

Tool: 'Foncier innovant' was developed in partnership with consulting firm Capgemini and Google. The tool used data captured by the government platform Le Géoportail and other available geographical information to locate pools and other outbuildings.

Role of the human: during its development, the project did not involve land surveyors, and their expertise and technical skills were not taken into account. Surveyors are now concerned about the gradual automation of their expertise, especially about the calculation of cadastral plans and the setting of the boundaries of private properties. Their lack of involvement means that the quality of the plans is not guaranteed, there may be an adverse impact on citizens and that the reliability of the whole process is limited. The public service that implements the AI tool has noticed that quality of service is decreasing.

Case 4 Serbian law on social cards

In 2022, the Serbian government introduced a social card system to promote administrative efficiency in the welfare system. This affected the Centres for Social Work across the country. The system centralised a government database of individuals who are recipients of or who apply for social security benefits, consolidating personal data from multiple data registries in a database that can be accessed by a significant number of employees in the social protection sector.

Tool: the system profiled individuals and used an automated decision-making tool able immediately and legally to suspend or reduce social benefits and social assistance, without considering people's life circumstances or allowing individuals to provide contextual or additional information (Amnesty International 2023; Government of the Republic of Serbia 2021).

Consequences: the system collected sensitive data, aggregating it from other databases (including information about ex partners) and violated privacy laws. In addition, 34,686 individuals lost their social benefits because their recorded earnings put them above the minimum threshold for assistance.

The NGO 'A11 Initiative for Economic and Social Rights' challenged the law establishing the social cards system before Serbia's Constitutional Court. The judgment is in preparation at the time of writing (A11 Initiative for Economic and Social Rights 2023).

Role of the human: social workers were unable to amend the data in the system or to override decision-making. The system uses a traffic light system with three options for social workers to click on: urgent, check, inform; and a 3-day deadline to perform a verification check in urgent notifications.

3. The precautionary principle – a legitimate way to address the risks of AI

In the EU context, the governance of AI will principally rely on the legislative tool that is the AI Act. Should the AI Act fail to deliver the necessary protection – by design or because the authorities involved either lack the resources to enforce it, act without the necessary level of coordination or because uncertainty is too high – there will be a need for solid legal principles to come to the rescue. The precautionary principle, which emanates from international environmental law, is such a principle. It represents an early warning system that 'enables decision-makers to adopt precautionary measures when scientific evidence about an environmental or human health hazard is uncertain and the stakes are high' (European Parliament 2015).

Developed in the early 1980s and formally adopted in 1992 at the UN Rio de Janeiro Conference on Environment and Development and in the UN Convention on Biodiversity, it was defined in 2005 by the UNESCO World Commission on the Ethics of Scientific Knowledge and Technology in the following manner:

> When human activities may lead to morally unacceptable harm that is scientifically plausible but uncertain, actions shall be taken to avoid or diminish that harm. Morally unacceptable harm refers to harm to humans or the environment that is threatening to human life or health, or serious and effectively irreversible, or inequitable to present or future generations, or imposed without adequate consideration of the human rights of those affected. (UNESCO 2005: 13)

In the EU, the principle was included in the Maastricht Treaty in 1992 and is there in Article 191 of the Treaty on the Functioning of the European Union. In practice, the precautionary principle has underpinned the EU's environmental policy and has been a core element of its risk and public health policies. In the area of chemicals policy, Article 1.3 of the Regulation on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) states that 'its provisions are underpinned by the precautionary principle'. In the field of EU food safety, Article 7 of the General Food Law Regulation states that 'when faced with these specific circumstances, decision makers or risk managers may take measures or other actions based on the precautionary principle'. The principle is also invoked in the fields of food safety and trade.

Further, as Mazur (2019) argues, some of its elements can be seen in the General Data Protection Regulation (GDPR): the right to be informed about the collection and use of personal data (Articles 13 and 14); and the right to assess the impact of such processing (Article 35).

The European Court of Justice has interpreted the precautionary principle in various cases, including the 1983 *Sandoz* case,[1] *Alpharma Inc*[2] and *Solvay Pharmaceuticals BV*.[3] On the latter, the Court expanded the scope of the principle from the protection of the environment to the protection of public health, concluding that 'in the domain of [human health], the existence of solid evidence which, while not resolving scientific uncertainty may reasonably raise doubts as to the safety of a substance justifies, in principle, [the refusal to include that substance…]. The precautionary principle is designed to prevent potential risks.'

While it is a well-established principle that can be legitimately invoked and applied, careful reflection should be given to how and when it should be applied to AI and how it interrelates with decision-making (Fisher et al. 2006; Stirling 2006; Donati 2021). Some criticise the principle for being paralysing and unscientific, and for promoting a culture of irrational fear; a reaction, as Aven (2023) argues, often rooted in a lack of clarity over the meaning and the scope of the principle. The EC, to address the controversies, has established guidelines for its application whereby it can be invoked only when three preliminary conditions are met: (a) identification of potentially adverse effects; (b) evaluation of the scientific data available; (c) the extent of scientific uncertainty (European Commission 2000).

If, as Hansson (2023) argues, precautionary actions are based on the current state of science, that potential dangers of limited plausibility are excluded and that the precautionary principle is not used to make judgments between competing top priorities, then it can be recognised as an essential principle that must be at the heart of technological development. With detailed procedures, standards and guidelines developed to enable its application in specific contexts (Aven 2023), the precautionary principle should formally be included in the governance of AI and other emerging technologies. It can sustain their development, give direction to innovation, help build a governance based on dialogue that involves relevant societal actors and contribute to ensuring that technological innovations are safe for society.

4. Human-in-control as a response to the risks of automation and AI

The human-in-control concept first appeared in the aviation sector with the introduction of automated aircraft control technology. Recognising that 'automation is able to limit

1. Case 174/82, *Sandoz BV* ECLI:EU:C:1983:213, para. 16. The CJEU recognised the idea underlying the precautionary principle by stating that 'in so far as there are uncertainties at the present state of scientific research it is for the Member States, in the absence of harmonization, to decide what degree of protection of the health and life of humans they intend to assure'. See also Guida (2021).
2. Case T-13/99 *Pfizer Animal Health SA v Council of the European Union* ECLI:EU:T:2002:209, para. 444: 'The institutions cannot be criticised for having chosen to withdraw provisionally the authorisation of virginiamycin as an additive in feedingstuffs, in order to prevent the risk from becoming a reality, and, at the same time, to continue with the research that was already under way. Such an approach, moreover, was consonant with the precautionary principle, by reason of which a public authority can be required to act even before any adverse effects have become apparent.'
3. Case T-392/02, *Solvay Pharmaceuticals BV v Council of the European Union* ECLI:EU:T:2003:277, para. 3.

the operator's authority' and that 'sometimes, it is not obvious for the operator to know that this has occurred', NASA designed principles for what it calls human-centred automation (Billings 1996).

The main axiom is that humans, in this instance the pilot and the air traffic controllers, bear the responsibility and remain in command: the first of their flights; the latter of air traffic more generally. The corollaries are that pilots and controllers: (a) must be actively involved; (b) must be adequately informed; (c) must be able to monitor the automation assisting them; (d) the automated systems must be predictable; (e) the automated systems must monitor the human operators; and (f) every intelligent system element must understand the intent of other intelligent system elements (Billings 1996).

An important additional remark is made: 'Though humans are far from perfect sensors, decision-makers and controllers, they possess three invaluable attributes. They are excellent detectors of signals in the midst of noise, they can reason effectively in the face of uncertainty, and they are capable of abstraction and conceptual organization' (Billings 1996).

Here, if we look at the world of work, a relevant parallel can be established with the role of workers' representatives who are present in the workplace and are reliable detectors of variables, weak signals, hazards and other factors that influence the work organisation and environment.

Human-in-control has also been adopted by the military, including in the use of AI-based military applications for planning, decision support and intelligence, in particular the Intelligence, Surveillance, Target Acquisition and Reconnaissance (ISTAR) capabilities developed for the armed forces.

The principle has been promoted both by international organisations and by national jurisdictions. The International Committee of the Red Cross has stressed the need for 'human control' of certain 'critical functions' of weapons systems, in particular their ability to 'select (search for, detect, identify, track or select) and attack (use force against, neutralize, damage or destroy) targets' without human intervention (Davison 2018).

In recent years, the principle has fully migrated to the AI sector and has been attributed with a range of different purposes by scholars (Davidovic 2023). Some qualify it as 'a key tool for assuring safety, dignity, and responsibility for AI and automated decision-systems' (Christen et al. 2023; Davidovic 2023). Other purposes relate to the need to ensure accuracy, safety and precision; to deliver accountability and responsibility; and to have sufficient understanding of the process to consent, dissent or to ensure trust in the institutions (Davidovic 2023).

As far as AI governance in the EU is concerned, no mention of human-in-control as such is made in the early EC communication on AI. Initially, the Commission referred to a 'human-centred' approach to AI and to the need to foster 'human-centric' digitalisation. The vision aims at ensuring people are at the centre and empowered, and that innovative business is fostered. It is with this vision in mind that, in December 2022, the European

Parliament, the Council and the Commission proclaimed a joint Declaration on Digital Rights and Principles for the Digital Decade, which they qualify as reflecting EU values and promoting a sustainable, human-centric vision for the digital transformation. Importantly, the Declaration should guide policymakers and put people at the centre of the digital transformation.

However, a shift of perspective has since taken place, from human-centric to human-in-control. Although not explicitly stated in the EC's digital agenda, human-in-control has gradually infused the speeches of high-level commissioners. When presenting the EC's digital strategy, Shaping Europe's Digital Future, President von der Leyen stated that 'Artificial intelligence must always comply with people's rights. This is why a person must always be in control of critical decisions' (President von der Leyen 2020). Later that same year, she also stated that 'we must ensure that our rights, privacy and protections are the same online as they are off it. That we can each have control over our own lives and over what happens to our personal information. That we can trust technology with what we say and do. That new tech does not come with new values' (von der Leyen 2020). In May 2023, in her speech at the 15th Congress of the ETUC, she stated that '[…] the answer to the challenges that AI raises is first of all a principle. This principle is called "human-in-control". That must be our underlying principle for everything' (von der Leyen 2023a). Then, in September 2023, in her speech at the Pulse Women Economic Network, she reaffirmed that 'the EU is promoting the "human-in-control" principle for sensitive applications of AI. Because the new digital world should not reproduce old inequalities, but open up new opportunities' (von der Leyen 2023b).

In 2017, in its opinion on 'Artificial Intelligence – the consequences of Artificial Intelligence on the (digital) single market, production, consumption, employment and society', the European Economic and Social Committee (EESC) had already called for humans to be in control of the technology, referring to a 'human-in-command' approach to AI 'including the precondition that the development of AI be responsible, safe and useful, where machines remain machines and people retain control over these machines at all times'. It called for 'transparent, comprehensible and monitorable AI systems, the operation of which is accountable, including retrospectively'. It also referred to the role of managers who should be involved 'so that they remain in control of these developments and are not the victims of them'. (European Economic and Social Committee 2017).

To operationalise human-in-control, such high-level recognition of its relevance and legitimacy in respect of AI governance is an essential prerequisite. As a concept, human-in-control goes beyond legislation and provides a supplementary level of protection, and it may well be needed if the obligations established under the AI Act and the standards attached to them become obsolete as the technology converges and becomes more intertwined in our lives.

However, it can only exist if humans are involved, not simply informed. In the world of work, while the EU legislation provides for information and consultation rights when

technology is introduced or modified in workplaces,4 the level of such information and consultation often have different interpretations. Here, one may cite the NASA principles as a source of useful inspiration as they establish that workers must play 'an active and necessary role apart from simply monitoring the course of the operation. That role may involve active control, or decision-making, or allocation of resources, or evaluation of alternatives, but it should not be passive, as it too often is today' (Billings 1996: 10).

In their contributions to this book, some of the authors address the various dimensions of human-in-control, including the need to consider the conditions of data production as well as the tools and equipment used to manufacture and market these systems, as Antonio A. Casilli argues. Benedetta Brevini reflects on the AI life cycle, the infrastructure and the often scarce resources it uses: should there be a discussion about who controls the essential infrastructures that power AI?

Real control also raises the question of adequate enforcement. Mario Guglielmetti insists on the correct allocation of competences between the sectoral competent authorities and the supervisory authorities. In their enforcement duties, these need to cooperate effectively and exchange relevant information, including at cross-border level.

As AI systems and other convergent technologies have the ability to influence humanity as a whole, in all its dimensions (physical and mental wellbeing, dignity and other fundamental rights), ensuring that humans are in control implies a multi-dimensional approach. Here, as Helga Nowotny has described, a movement devoted to digital humanism has appeared and is attempting to integrate a human-centred approach in the design, production and deployment of AI throughout their systemic interlinkages (Nowotny 2021). This seeks to identify specific intervention points and to be attentive to actual practices in various domains, as well as become part of the education system. The values on which digital humanism is based will be crucial for shaping the future of work and of liberal democratic societies.

The deployment of AI in the workplace could therefore support the improvement of working conditions and the prevention of occupational risks, but it can also exacerbate the deterioration of working conditions if the tools provided reinforce strategic and organisational orientations that endanger workers' physical and mental health. In this context, increasing AI literacy is indispensable. As the degree to which humans can exercise meaningful control is essential (Cavalcante Siebert et al. 2023), workers can increase their level of control if they become critical agents, able to understand the role of AI at work and its impact on their occupation, and anticipate how it may transform their careers, skills and roles. Passively using AI systems does not benefit them: a certain distance is needed for workers to see AI's overall influence (Ponce del Castillo 2020, 2023). They would need to be able to distinguish situations in which AI systems are effective or not (Brynjolfsson et al. 2023) and be sure whether they can use the technology reliably. This implies a shared understanding of technological advance and its consequences for work, as María Luz Rodríguez Fernández argues in these pages.

4. See Annex I Point 1(a) of the EU European Works Council Directive 2009/38.

Explicit discussions about the responsibilities and liabilities of each actor, including mechanisms for overruling the AI system 'through intervening and correcting behavior, setting new goals, or delegating sub-tasks' (Cavalcante Siebert et al. 2023) must also be had. This further entails an understanding of the costs behind the production of AI, in particular the natural resources used in its production, deployment and maintenance. Workers and trade unions need to develop this new skill which can help them navigate volatile and fast-moving technological developments.

As German Bender explains in his chapter, trade unions and employers need to be able to bargain – or codetermine – AI systems, regulating their use conditions and their possible known and unknown effects. This would extend the possibility of worker participation to areas usually beyond their reach because they are reserved to corporate actors, including 'black box' algorithms. The right to meaningful participation should, in the view of Rodríguez Fernández, allow worker representatives to 'see inside' and utilise what they see to guarantee that the decisions do not cause bias or differentiated treatment without justification.

Looking beyond the world of work, respect for intellectual and emotional autonomy is another essential dimension, as Frank Pasquale argues. Genuine control also requires an inclusive and participatory governance, involving a wide range of stakeholders including from marginalised and disadvantaged groups (see Ulnicane, this volume).

Finally, Helga Nowotny contends that exercising genuine control entails a reassessment of profit allocation. The inequitable distribution of the productivity gains stemming from AI systems raises the question of whether those responsible for generating profit for the technology sector should also benefit from it. This encompasses the workers involved in the development and application of AI, as well as those whose data is utilised to improve or generate the technology (Tubaro et al. 2020). It may therefore be necessary to reconsider how workers can participate not just in AI extraction and production, but also in the ensuing economic rewards for the data they provide (Brynjolfsson et al. 2023), possibly in the form of wage increases. This volume does not address this matter directly, but it does approach it indirectly.

5. Conclusion

As illustrated by the four case studies presented here, the risks posed by AI are serious and are having an impact on the lives of thousands of people globally. To address those risks better and to prevent harm, the point of initial governance must be moved to an earlier stage of technological development. This implies giving the legal precaution principle and the concept of human-in-control a central role in the governance approach. Such a proactive, anticipative and collaborative form of precautionary governance can improve our prospects of safeguarding society from AI's potential impacts. With the increasing autonomy of AI systems and the recent development of generative AI, the risk of creating AI systems that pursue undesirable goals becomes real and the need for control and effective human intervention even greater (Bengio et al. 2023).

The need for human control is one of the perspectives discussed in this book, which presents the insights of prominent scholars and specialists in the field hailing from Europe and other parts of the globe. When the project was initiated, the intention was to assess the challenges that AI poses, pinpoint the crucial aspects of a potential response and highlight some possible fundamental constituents of an all-encompassing framework for AI governance. We modestly hope the book achieves some of these objectives.

References

A11 Initiative for Economic and Social Rights (2023) Support grows for A 11 constitutional challenge to the social cards law. https://www.a11initiative.org/en/support-grows-for-a-11-constitutional-challenge-to-the-social-cards-law/

Amnesty International (2023) Serbia submission for European Union enlargement package/opinion, 2023. https://www.amnesty.org/en/wp-content/uploads/2023/04/EUR7066882023ENGLISH.pdf

Aven T. (2023) A risk and safety science perspective on the precautionary principle, Safety Science, 165, 106211. https://doi.org/10.1016/j.ssci.2023.106211

Billings C.E. (1996) Human-centered aviation automation: Principles and guidelines, NASA Technical Memorandum 110381, National Aeronautics and Space Administration.

Bengio Y. et al. (2023) Managing AI risks in an era of rapid progress. https://managing-ai-risks.com/

Brynjolfsson E., Li D. and Raymond L.R. (2023) Generative AI at work, Working Paper 31161, National Bureau of Economic Research. https://doi.org/10.3386/w31161

Cavalcante Siebert L. et al. (2023) Meaningful human control: Actionable properties for AI system development, AI and Ethics, 3 (1), 241–255. https://doi.org/10.1007/s43681-022-00167-3

Christen M., Burri T., Kandul S. and Vörös P. (2023) Who is controlling whom? Reframing 'meaningful human control' of AI systems in security, Ethics and Information Technology, 25 (1), 10. https://doi.org/10.1007/s10676-023-09686-x

Collingridge D. (1982) The social control of technology, Palgrave Macmillan.

Cortez N. (2019) Digital health and regulatory experimentation at the FDA, Yale Journal of Law and Technology, 21 (4), 4–26.

Coughlan S. (2020) Scottish school pupils have results upgraded, BBC News, 8 November 2020. https://www.bbc.co.uk/news/uk-scotland-53740588.amp

Davidovic J. (2023) On the purpose of meaningful human control of AI, Frontiers in big data, 5. https://doi.org/10.3389/fdata.2022.1017677

Davison N. (2018) A legal perspective: Autonomous weapon systems under international humanitarian law. https://www.icrc.org/en/download/file/65762/autonomous_weapon_systems_under_international_humanitarian_law.pdf

Direction générale des Finances publiques (2022) L'intelligence artificielle au service de la lutte contre la fraude : bilan de l'expérimentation « foncier innovant ». https://www.impots.gouv.fr/sites/default/files/media/2_actu/home/2022/dp_foncier_innovant.pdf

Donati A. (2021) The precautionary principle under European Union law, Hitotsubashi Journal of Law and Politics, 49, 43–60. https://doi.org/10.15057/hjlp.2020003

European Commission (2000) Communication from the Commission on the precautionary principle, COM(2000) 1 final, 2.2.2000. https://op.europa.eu/en/publication-detail/-/publication/21676661-a79f-4153-b984-aeb28f07c80a/language-en

European Commission (2021) Proposal for a regulation of the European Parliament and of the Council on machinery products, COM(2021) 202 final, 21.4.2021. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=COM:2021:202:FIN

European Economic and Social Committee (2017) Opinion of the European Economic and Social Committee on: Artificial intelligence – The consequences of artificial intelligence on the (digital) single market, production, consumption, employment and society, Official Journal of the European Union, C 288, 31.8.2017. https://www.eesc.europa.eu/en/our-work/opinions-information-reports/opinions/artificial-intelligence-consequences-artificial-intelligence-digital-single-market-production-consumption-employment-and

European Parliament (2015) The precautionary principle: Definitions, applications and governance, European Parliamentary Research Service. https://www.europarl.europa.eu/RegData/etudes/IDAN/2015/573876/EPRS_IDA(2015)573876_EN.pdf

European Parliament (2022) The Dutch childcare benefit scandal, institutional racism and algorithms. https://www.europarl.europa.eu/doceo/document/O-9-2022-000028_EN.html

Fisher E.C., Jones J.S. and von Schomberg J.R. (2006) Implementing the precautionary principle: Perspectives and prospects, in Fisher E.C., Jones J.S. and von Schomberg R. (eds.) Implementing the precautionary principle: Perspectives and prospects, Edward Elgar, 1–18.

Government of the Republic of Serbia (2021) Government passes social card bill. https://www.srbija.gov.rs/vest/en/166629/government-passes-social-card-bill.php

Guida A. (2021) The precautionary principle and genetically modified organisms: A bone of contention between European institutions and Member States, Journal of Law and the Biosciences, 8 (1). https://doi.org/10.1093/jlb/lsab012

Hadwick D. and Lan S. (2021) Lessons to be learned from the Dutch childcare allowance scandal: A comparative review of algorithmic governance by tax administrations in the Netherlands, France and Germany, World Tax Journal, 13 (4), 609–645.

Hansson S.O. (2020) How extreme is the precautionary principle?, Nanoethics, 14 (3), 245–257. https://doi.org/10.1007/s11569-020-00373-5

Henley J. (2021) Dutch government faces collapse over child benefits scandal, The Guardian, 14 January 2021. https://www.theguardian.com/world/2021/jan/14/dutch-government-faces-collapse-over-child-benefits-scandal

Kaal W.A. (2016) Dynamic regulation for innovation, in Fenwick M., Kaal W.A., Kono T. and Vermeulen E.P.M. (eds.) Perspectives in law, business and innovation, Springer, 16–22.

Kippin S. and Cairney P. (2022) The Covid-19 exams fiasco across the UK: Four nations and two windows of opportunity, British Politics, 17 (1), 1–23. https://doi.org/10.1057/s41293-021-00162-y

Kolkman D. (2020) F** k the algorithm?: What the world can learn from the UK's A-level grading fiasco, Impact of Social Sciences Blog, 26 August 2020. https://eprints.lse.ac.uk/106366/1/impactofsocialsciences_2020_08_26_fk_the_algorithm_what_the_world_can.pdf

Linkov I. et al. (2018) Comparative, collaborative, and integrative risk governance for emerging technologies, Environment Systems and Decisions, 38 (2), 170–176. https://doi.org/10.1007/s10669-018-9686-5

Mandel G.N. (2013) Emerging technology governance, in Marchant G.E., Abbott K.W. and Brown J.E. (eds.) Innovative governance models for emerging technologies, Edward Elgar, 44–62.

Mandel G.N. (2020) Regulating emerging technologies, in Marchant G.E. and Wallach W. (eds.) Emerging technologies: Ethics, law and governance, Routledge.

Mazur J. (2019) Automated decision-making and the precautionary principle in EU law, TalTech Journal of European Studies, 9 (4), 3–18. https://doi.org/10.1515/bjes-2019-0035

Mieg H.A. (ed.) (2022) The responsibility of science, Springer. https://doi.org/10.1007/978-3-030-91597-1

Nowotny H. (2021) In AI we trust: Power, illusion and control of predictive algorithms, John Wiley & Sons.

Office for Artificial Intelligence (2020) A guide to using artificial intelligence in the public sector. https://www.gov.uk/government/publications/a-guide-to-using-artificial-intelligence-in-the-public-sector

Ponce Del Castillo A. (2020) Labour in the age of AI: Why regulation is needed to protect workers, Foresight Brief 08, ETUI. https://www.etui.org/publications/foresight-briefs/labour-in-the-age-of-ai-why-regulation-is-needed-to-protect-workers

Ponce Del Castillo A. (2023) AI: Discovering the many faces of a faceless technology: A hands-on tool to help map AI, strengthen critical thinking and support anyone involved in negotiating the deployment of AI systems, ETUI. https://www.etui.org/publications/ai-discovering-many-faces-faceless-technology-0

Pouget H. and Laux J. (2023) A letter to the EU's future AI office, Carnegie Endowment for International Peace, 3 October 2023. https://carnegieendowment.org/2023/10/03/letter-to-eu-s-future-ai-office-pub-90683

Robbins S. (2023) The many meanings of meaningful human control, AI and Ethics, 1–12. https://doi.org/10.1007/s43681-023-00320-6

Rotolo D., Hicks D. and Martin B.R. (2015) What is an emerging technology?, Research policy, 44 (10), 1827–1843. https://doi.org/10.1016/j.respol.2015.06.006

Schwarz E. (2018) The (im)possibility of meaningful human control for lethal autonomous weapon systems, Humanitarian law and policy, 29 August 2018. https://blogs.icrc.org/law-and-policy/2018/08/29/im-possibility-meaningful-human-control-lethal-autonomous-weapon-systems/

Stirling A. (2006) Precaution, foresight and sustainability: Reflection and reflexivity in the governance of science and technology, in Voss J. and Kemp R. (eds.) Reflexive governance for sustainable development, Edward Elgar, 225–272.

Taeihagh A., Ramesh M. and Howlett M. (2021) Assessing the regulatory challenges of emerging disruptive technologies, Regulation and Governance, 15 (4), 1009-1019. https://doi.org/10.1111/rego.12392

Ten Seldam B. and Brenninkmeijer A. (2021) The Dutch benefits scandal: A cautionary tale for algorithmic enforcement, EU Law Enforcement, 30 April 2021. https://eulawenforcement.com/?p=7941

Tegtmeier P., Weber C., Sommer S., Tisch A. and Wischniewski S. (2022) Criteria and guidelines for human-centered work design in a digitally transformed world of work: Findings from a formal consensus process, International Journal of Environmental Research and Public Health, 19 (23), 15506. https://doi.org/10.3390/ijerph192315506

Tubaro P., Casilli A.A. and Coville M. (2020) The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence, Big Data and Society, 7 (1). https://doi.org/10.1177/2053951720919776

UNESCO (2005) The precautionary principle, World Commission on the Ethics of Scientific Knowledge and Technology. https://unesdoc.unesco.org/ark:/48223/pf0000139578

van Bekkum M. and Borgesius F.Z. (2021) Digital welfare fraud detection and the Dutch SyRI judgment, European Journal of Social Security, 23 (4), 323–340. https://doi.org/10.1177/13882627211031257

von der Leyen (2020a) Press remarks by President von der Leyen on the Commission's new strategy: Shaping Europe's digital future, 19 February 2020. https://ec.europa.eu/commission/presscorner/detail/nl/speech_20_294

von der Leyen (2020b) Shaping Europe's digital future: Op-ed by Ursula von der Leyen, President of the European Commission, 19 February 2020. https://ec.europa.eu/commission/presscorner/detail/es/ac_20_260

von der Leyen (2023a) Speech by President von der Leyen at the 15th Congress of the European Trade Union Confederation, 25 May 2023. https://ec.europa.eu/commission/presscorner/detail/en/speech_23_2926

von der Leyen (2023b) Speech by President von der Leyen at the Pulse Women Economic Network, via video message, 5 September 2023. https://ec.europa.eu/commission/presscorner/detail/en/speech_23_4351

Zhang B. Anderljung M., Kahn L., Dreksler N., Horowitz M.C. and Dafoe A. (2021) Ethics and governance of artificial intelligence: Evidence from a survey of machine learning researchers, Journal of Artificial Intelligence Research, 71, 591-666. https://doi.org/10.48550/arXiv.2105.02117

Zuboff S. (2019) The age of surveillance capitalism: The fight for a human future at the new frontier of power, PublicAffairs.

All links were checked on 19.01.2023.

Cite this chapter: Ponce del Castillo A. (2024) AI: the value of precaution and the need for human control, in Ponce del Castillo A. (ed.) Artificial intelligence, labour and society, ETUI.