ΠΑΝΤΕΙΟΝ ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΟΙΝΩΝΙΚΩΝ ΚΑΙ ΠΟΛΙΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

PANTEION UNIVERSITY OF SOCIAL AND POLITICAL SCIENCES

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΟΙΚΟΝΟΜΙΑΣ & ΔΗΜΟΣΙΑΣ ΔΙΟΙΚΗΣΗΣ

ΤΜΗΜΑ ΔΗΜΟΣΙΑΣ ΔΙΟΙΚΗΣΗΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΔΗΜΟΣΙΑ ΔΙΟΙΚΗΣΗ»

ΚΑΤΕΥΘΥΝΣΗ: ΔΙΚΑΙΟ, ΤΕΧΝΟΛΟΓΙΑ ΚΑΙ ΟΙΚΟΝΟΜΙΑ

The Ethical and Legal Challenges of Artificial Intelligence:

The EU response to biased and discriminatory AI

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αναστασία Σιάπκα (Α.Μ. 7116Μ106)

Αθήνα, 2018

Τριμελής Επιτροπή

Αντώνιος Χάνος, Αναπληρωτής Καθηγητής (Επιβλέπων)
Ελευθέριος Βόγκλης, Επίκουρος Καθηγητής
Φερενίκη Παναγοπούλου, Επίκουρη Καθηγήτρια

The Ethical and Legal Challenges of Artificial Intelligence: The EU response to biased and discriminatory AI

Abstract

The proliferation of Artificial Intelligence (AI) in decision-making contexts is hailed as a silver bullet, pledging to replace human subjectivity with objective, infallible decisions. Paradoxically, considerable journalistic reporting has recently commanded attention to biased and discriminatory attitudes displayed by AI systems on both sides of the Atlantic. Notwithstanding the permeation of automated decision-making in critical settings, such as criminal justice, job recruitment, and border control, wherein rights and freedoms of individuals and groups are likewise imperilled, there is often no way for human agents to untangle how AI systems reach such unacceptable decisions. The conspicuous bias problem of AI alongside its operation as an inexplicable 'black box' render the exploration of this phenomenon pressing, primarily in the less examined EU policy arena.

This dissertation pursues an interdisciplinary research methodology to examine which are the main ethical and legal challenges that Narrow AI, especially in its data-driven Machine Learning (ML) form, poses in relation to bias and discrimination across the EU.

Chapter 1 equips readers with pertinent background information regarding AI and its interdependent ML and Big Data technologies. In an accessible manner, it takes heed of the definitions and types of AI adopted by EU instruments along with the milestones in its historical progression and its current stage of development.

Chapter 2 conducts a philosophical analysis to argue against the putative ethical neutrality of AI. Ethical concerns of epistemological nature reveal that biases traverse AI systems through the selection of objectives, training data, the reliance on correlations, and the epistemic inequality between lay individuals and AI developers in combination with that between human agents and 'black box' machines in general. Touching upon normative ethical concerns, AI systems entail effects which, according to egalitarianism, oppose normative ideals of fairness and equality. In more Kafkaesque scenarios, individuals and corporations may use technical particularities of AI to mask their discriminatory intent.

In Chapter 3, a doctrinal legal methodology is applied to reveal the tensions of these challenging instantiations of AI in light of soft and hard EU law instruments. In consideration of its data-driven character, biased and discriminatory AI decisions fall within the applicability scope of the newly enforced General Data Protection Regulation (GDPR). In particular, the data processing principles of Article 5, the Data Protection Impact Assessments (DPIA) of Article 35, the prohibition of automated decision-making and the speculative right to explanation of Article 22, the principles of lawfulness, fairness, and transparency of Article 5

(1) a), the suggested implementation of auditing, and the enhanced enforcement authorities receive scrutiny.

The dissertation concludes that a principles-based approach and the provision of anticipatory impact assessments are regulatory strengths of the GDPR. However, the EU should discourage the deployment of AI in crucial decision-making contexts and explore ways to fill related legal gaps. Overall, Trustworthy AI is proposed as an ethical and legal paragon in the face of biased and discriminatory AI.

Keywords: artificial intelligence; discrimination; machine learning; automated decision-making; General Data Protection Regulation; bias

Περίληψη

Η ταχεία διάδοση της Τεχνητής Νοημοσύνης (ΤΝ) στα πλαίσια λήψης αποφάσεων τυγχάνει ενθουσιώδους υποδοχής ως λύση υποσχόμενη να αντικαταστήσει την ανθρώπινη υποκειμενικότητα με αντικειμενικές, αλάνθαστες αποφάσεις. Παραδόξως, αξιόλογες δημοσιογραφικές έρευνες έστρεψαν πρόσφατα την προσοχή σε περιπτώσεις όπου συστήματα ΤΝ και στις δύο πλευρές του Ατλαντικού επέδειξαν μεροληπτικές στάσεις. Παρά τη διάχυση της αυτοματοποιημένης λήψης αποφάσεων σε ύψιστης σημασίας διαδικασίες, όπως κατά την απονομή ποινικής δικαιοσύνης, την εύρεση εργασίας και τον έλεγχο των συνόρων, όπου τα δικαίωματα και οι ελευθερίες ατόμων και κοινωνικών ομάδων διακυβεύονται εξίσου, δεν υπάρχει μέθοδος που επιτρέπει στους ανθρώπους να αποσαφηνίσουν πώς η Τεχνητή Νοημοσύνη καταλήγει σε τέτοιες απαράδεκτες αποφάσεις. Το ευδιάκριτο πρόβλημα μεροληψίας της Τεχνητής Νοημοσύνης σε συνδυασμό με τη λειτουργία της ως ανεξήγητο «μαύρο κουτί» καθιστούν την εξερεύνηση αυτού του φαινομένου επιτακτική, πρωτίστως στο λιγότερο εξετασθέν πεδίο των Ενωσιακών πολιτικών.

Η παρούσα διπλωματική εργασία επιδιώκει μια διεπιστημονική ερευνητική μεθοδολογία για να εξετάσει ποιες είναι οι κύριες ηθικές και νομικές προκλήσεις που θέτει η στενή Τεχνητή Νοημοσύνη (Narrow AI), ειδικά στη μορφή της οδηγούμενης από δεδομένα (data-driven) Μηχανικής Μαθήσεως (Machine Learning), σε θέματα διακρίσεων και προκαταλήψεων στην ΕΕ.

Το Κεφάλαιο 1 εφοδιάζει τους αναγνώστες με το τεχνικό υπόβαθρο αναφορικά με την Τεχνητή Νοημοσύνη και τις αλληλένδετες τεχνολογίες Μηχανικής Μαθήσεως (Machine Learning) και Μεγάλων Δεδομένων (Big Data). Με προσιτό τρόπο, εστιάζει στους ορισμούς και τύπους Τεχνητής Νοημοσύνης που υιοθετούνται σε Ενωσιακές πράξεις καθώς και στα ορόσημα της ιστορικής της εξέλιξης και το τρέχον στάδιο ανάπτυξής της.

Το Κεφάλαιο 2 διεξάγει μια φιλοσοφική ανάλυση για να επιχειρηματολογήσει κατά της υποτιθέμενης ηθικής ουδετερότητας της Τεχνητής Νοημοσύνης. Ηθικοί προβληματισμοί επιστημολογικής φύσεως αποκαλύπτουν πως οι προκαταλήψεις διασχίζουν τα συστήματα ΤΝ μέσω της επιλογής στόχων, των δεδομένων εκπαίδευσης (training data), της εξάρτησης από συσχετισμούς και της επιστημολογικής ανισότητας μεταξύ μη ειδημόνων και προγραμματιστών ΤΝ καθώς και μεταξύ ανθρώπων και μηχανών τύπου «μαύρου κουτιού» γενικότερα. Περνώντας σε κανονιστικά ηθικά ζητήματα, τα συστήματα ΤΝ επιφέρουν επιπτώσεις που, σύμφωνα με την εξισωτική θεωρία, αντιτίθενται στα κανονιστικά ιδανικά της δικαιοσύνης και της ισότητας. Σε περισσότερο καφκικά σενάρια, άτομα και εταιρείες μπορούν

να χρησιμοποιήσουν τις τεχνικές ιδιαιτερότητες της Τεχνητής Νοημοσύνης προκειμένου να κρύψουν μεροληπτικές προθέσεις.

Στο Κεφάλαιο 3 εφαρμόζεται η δογματική νομική μεθοδολογία για να αποκαλύψει τις εντάσεις αυτών των προβληματικών εκδηλώσεων της ΤΝ υπό το πρίσμα του ήπιου και αυστηρού Ενωσιακού δικαίου. Λαμβάνοντας υπόψιν τον χαρακτήρα της ως οδηγούμενης από δεδομένα (data-driven), οι μεροληπτικές αποφάσεις της ΤΝ εμπίπτουν στο πεδίο εφαρμογής του νέου Γενικού Κανονισμού Προστασίας Δεδομένων (ΓΚΠΔ, General Data Protection Regulation). Ειδικότερα, διερευνώνται οι αρχές που διέπουν την επεξεργασία δεδομένων κατά το Άρθρο 5, οι εκτιμήσεις αντικτύπου σχετικά με την προστασία δεδομένων (ΕΑΠΔ, DPIA) του Άρθρου 35, η απαγόρευση της αυτοματοποιημένης λήψης αποφάσεων και το εικαζόμενο δικαίωμα αιτιολόγησης του Άρθρου 22, οι αρχές της νομιμότητας, της δικαιοσύνης και της διαφάνειας στο Άρθρο 5 (1) α), η προτεινόμενη επιβολή ελέγχων και οι ενισχυμένες αρχές προστασίας δεδομένων.

Η διπλωματική εργασία συμπεραίνει ότι η βασισμένη σε αρχές προσέγγιση καθώς και οι εκτιμήσεις αντίκτυπου σχετικά με την προστασία δεδομένων αποτελούν ισχυρά ρυθμιστικά σημεία του ΓΚΠΔ. Ωστόσο, η ΕΕ πρέπει να αποθαρρύνει την επέκταση της Τεχνητής Νοημοσύνης σε κρίσιμα πεδία λήψης αποφάσεων και να διερευνήσει τρόπους κάλυψης των σχετικών νομικών κενών. Σε γενικές γραμμές, η «Αξιόπιστη ΤΝ» (Trustworthy AI) προτείνεται ως ηθικό και νομικό πρότυπο εν όψει της προκατειλημμένης Τεχνητής Νοημοσύνης που επιδίδεται σε διακριτική μεταχείριση.

Λέξεις κλειδιά: τεχνητή νοημοσύνη, διακριτική μεταχείριση, μηχανική μάθηση, αυτοματοποιημένη λήψη αποφάσεων, γενικός κανονισμός προστασίας δεδομένων, προκαταλήψεις

Table of Contents

List of abbreviations

AI	Artificial Intelligence
AI HLEG	High-Level Expert Group on Artificial Intelligence
AIA(s)	Algorithmic Impact Assessment(s)
AJL	Algorithmic Justice League
CFREU	Charter of Fundamental Rights of the European Union
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
DARPA	Defense Advanced Research Projects Agency
DL	Deep Learning
DPIA(s)	Data Protection Impact Assessment(s)
EESC	European Economic and Social Committee
EGE	European Group on Ethics in Science and New Technologies
EPRS	European Parliamentary Research Service
EPSC	European Political Strategy Centre
EU	European Union
FAT-ML	Fairness, Accountability, and Transparency in Machine Learning
FRA	European Union Agency for Fundamental Rights
GDPR	General Data Protection Regulation
HRESIA	Human Rights, Ethical and Social Impact Assessment
IBM	International Business Machines Corporation
ICO	Information Commissioner's Office
MIT	Massachusetts Institute of Technology
ML	Machine Learning
STEM	Science, Technology, Engineering, Mathematics
TEU	Treaty on European Union
TFEU	Treaty on the Functioning of the European Union
UK	United Kingdom
US	United States
WP29	Article 29 Data Protection Working Party
XAI	Explainable Artificial Intelligence

Introduction

Background

From policy-makers and academics to industry leaders and civil society organisations, a heated debate has sparked regarding the regulation of Artificial Intelligence (AI), Machine Learning (ML), and Big Data. In spite of the positive impact that these emerging technologies bear on humanity, they increasingly seem to be characterised by vulnerabilities when processing individuals' data. Just as every creation, AI and its encompassing technologies carry with them and perpetuate a major flaw of their human creators: biases. The following incidents, which recently made headlines, are illustrative of this.

In an attempt to ensure objective legal procedures, the US has adopted AI tools that predict recidivism. In May 2016, the newsroom ProPublica published its investigative report regarding one such widely used AI tool, the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), created by the company Northpointe.[1] According to the report, the scores assigned by COMPAS were not only false in predicting criminal behaviour, but also discriminatory against black defendants. The AI algorithm wrongly predicted black defendants as twice more likely to commit crimes than white ones.[2] In parallel, it mislabelled white defendants as low-risk more frequently than black ones.[3] The ways in which the AI algorithm reached these predictions are unknown to the public and the defendants themselves, because COMPAS constitutes private property of Northpointe; as such, its functions are undisclosed.[4]

In the same year, Microsoft's conversational AI (chatbot), Tay, made her debut.[5] Tay ran on ML to hold sophisticated conversations on Twitter: the more people would converse with her, the better and more specific her responses would be. As soon as online trolls understood how Tay worked, they inundated her with hateful content. Shortly, she had integrated this content and was publicly hurling her own racist, anti-Semitic, sexist hate speech back to Twitter users. In less than 24 hours after her launch, Microsoft shut Tay down.[6]

---

[1] Julia Angwin et al., 'Machine Bias', ProPublica, 23 May 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[2] Angwin et al.
[3] Angwin et al.
[4] Angwin et al.
[5] Jonathan Vanian, 'Unmasking A.I.'s Bias Problem', Fortune, 25 June 2018, http://fortune.com/longform/ai-bias-problem/.
[6] Vanian.

In 2018, Google announced that it had fixed its Photos application, which was mistakenly identifying black people as gorillas.[7] The application uses ML to recognise people, places, and events depicted in photographs and automatically group those with similar content. Three years after an African American consumer and developer pointed out that Google Photos labelled photographs of his friends and him with the tag 'gorillas', all that the company managed to fix was completely removing the tag so that the ML algorithm does not assign it to any image whatsoever. According to the user who reported the incident, more diversity in the developers' team or testing to a diverse focus group would have helped the company identify the error and fix the application before its market launch.[8]

In October of the same year, Reuters reported that Amazon ceased the recruiting tool it was developing for the last quadrennium.[9] Using AI, specifically ML, the tool reviewed job applicants' résumés and assigned to each of them a score from one to five stars. Fortunately no later than its experimental stage, the company discovered that the AI tool was discriminating against women.[10] To produce its ranking, the AI was trained by spotting patterns in résumés submitted to the company during the last decade. However, given the broad underrepresentation of female professionals in the technology industry, the vast majority of these résumés belonged to male candidates. Based upon these data, the AI system learned to favour males.[11] It unfavourably graded résumés which included the word 'women's' (e.g. women's chess club captain) or the names of all-women's colleges, whereas it rewarded those including verbs such as 'executed' and 'captured', commonly found in male engineers' self-descriptions.[12] Despite Amazon's efforts to make the AI system neutral towards these terms, the company admitted its inability to guarantee that the system would not discriminate based on other variables. Eventually, it dismantled the development team and reassured the public that its recruiters had refrained from depending their evaluations on it.[13]

In November 2018, media reported that the EU will launch trials of an AI lie detector for border control, called iBorderCtrl, at checkpoints in Hungary, Greece, and Latvia. By

---

[7] Tom Simonite, 'When It Comes to Gorillas, Google Photos Remains Blind', *Wired*, 11 January 2018, https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/.
[8] Taryn Finley, 'Google Apologizes for Tagging Photos of Black People as "Gorillas"', HuffPost UK, 2 July 2015, http://www.huffingtonpost.com/2015/07/02/google-black-people-goril_n_7717008.html.
[9] Jeffrey Dastin, 'Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women', *Reuters*, 10 October 2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.
[10] Dastin.
[11] In this dissertation data are referred to in plural: 'Definition of Data', Oxford Dictionaries | English, accessed 12 November 2018, https://en.oxforddictionaries.com/definition/data.
[12] Dastin, 'Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women'.
[13] Dastin.

analysing travellers' micro-expressions, the AI system, which uses ML techniques, is reportedly able to detect whether they are lying. In response to the announcement, academics from the University of Amsterdam and University College London underlined that there is no scientific foundation for the methods of such systems, which will translate to unfair outcomes for individuals.[14] Some of them referred to these systems as facilitating a *'pseudoscientific border control'*, whereas others consider the use of such opaque tools to evaluate and classify people as a *'terrible idea'*.[15]

This selection of incidents shows that both in the US and recently in the EU data-driven AI, especially of the ML kind, faces a serious 'bias problem'. The pervasiveness of AI in decision procedures, even in crucial sectors such as policing, employment, and border control, means that harms caused to humans by biased data processing necessitate immediate action. The risk of such unacceptable processing looms no less for prominent private (Microsoft, Google, Amazon) and public entities (US judicial branch, EU). Despite the EU's anti-discrimination acquis, AI biases do not fit comfortably within its framework. Most of the times, biased AI emerges as a spin-off unintended, and even further unpredicted, by AI developers. Critically, the most useful AI systems function as 'black boxes', because of the human inability to explain how their algorithms turn their inputs to conclusions. Apart from an epistemic and normative hurdle, this inexplicability of AI incurs legal concerns, as it makes it impossible to explain the causal connection and underlying reasoning of algorithmic discrimination, as opposed to commonplace human discrimination, and to thereby build a legal case. All in all, the technical uncertainty of AI systems as indecipherable black boxes exhibiting biases elicits, in turn, ethical and legal uncertainty, which renders policy responses to AI a thorny affair.

Research question and outline

Intrigued by such incidents, this dissertation sets out to answer the following research question:

> ***Which are the main ethical and legal challenges that Narrow AI, especially in its data-driven ML form, poses in relation to bias and discrimination at the EU level?***

To address this question, the dissertation is divided into three sections. In Chapter 1, it provides a technical overview consisting of the definition of AI alongside its types, historical

[14] Daniel Boffey, 'EU Border "lie Detector" System Criticised as Pseudoscience', *The Guardian*, 2 November 2018, sec. World news, https://www.theguardian.com/world/2018/nov/02/eu-border-lie-detection-system-criticised-as-pseudoscience.
[15] Boffey; Rob Picheta, 'Passengers to Face AI Lie Detectors at EU Airports', CNN Travel, 1 November 2018, https://www.cnn.com/travel/article/ai-lie-detector-eu-airports-scli-intl/index.html.

development, and relation to ML and Big Data. To prepare the ground for a scientifically informed analysis in the successive Chapters, the focus of the dissertation is pinned down on AI with the following characteristics: Narrow AI systems, belonging to the family of ML techniques, grounded in (Big) Data. In Chapter 2, the dissertation argues that, in contrast to a widespread view of AI as objective, its ethical neutrality is compromised because of epistemic and normative shortcomings. Chapter 3 hinges on the abovementioned ethical concerns to explicate their conversion to legal ramifications and argue that AI is in tension with EU data protection law. Finally, the EU framework, as conceptualised through the ethical and legal argumentation, is evaluated and situated within recent initiatives. The dissertation concludes with an attempt to furnish pointers for further research.

Research limitations

First of all, the positive impacts of AI are uncontested but mentioned only in passing, as an examination of both positive and challenging aspects of AI would by far exceed the scope of a dissertation. Focusing on AI challenges which are of policy interest to the EU, the themes under elaboration are ethical and legal ones. Notwithstanding the significance of economic, social, environmental, and research implications of AI, they fall outside the limited ambit of this study. Yet, as these areas could benefit from an ethical and legal blueprint, the herein presented analysis is incidentally of relevance to them as well.

The legal and ethical challenges of AI weigh heavily upon multifaceted legal fields, such as liability, international humanitarian law, intellectual property, or data protection, and ethical enquiries, such as governance, ethical research, fairness, or transparency. To avoid a peripheral review of all these and allow for an in-depth, critical elaboration, the dissertation centres itself on the ethical and legal aspects of biased and discriminatory data processing with the use of Narrow AI, in public and private domains indiscriminately.

Moreover, the interdependence of AI, ML, and (Big) Data makes it impossible for a comprehensive analysis to examine these technologies separately. Thus, when scrutinising the legal and ethical challenges, this dissertation presumes the ways in which these technologies work with and empower each other as well as their synergistic effects.

A final limitation on examining and harnessing the legal ramifications of biased AI is the absence of relevant EU case-law to date, as these issues are recent in the regulatory as opposed to the academic and industry realms. Hence, the dissertation will harken back to the introductory examples of AI systems when in need to flesh out the analysis with factual context.

Research methodology

This dissertation adopts an interdisciplinary desk-research methodology, in the sense that it brings together two distinct approaches. In Chapter 2, referring to the ethical challenges, it follows the methodology of philosophical analysis, specifically from an epistemological and normative perspective. In Chapter 3, referring to the legal challenges, it follows a problem-based doctrinal (known as 'traditional legal' in Civil Law or 'black-letter' in Common Law traditions) methodology, which enables the gathering, description, and deductive application of EU legal rules to biased and discriminatory AI. In trying to answer 'what is the law?' in the case of biased AI, the doctrinal research concentrates on sections of EU law and legal literature but is likewise supplemented by non-binding soft law considerations, in order to deduce both de lege lata and de lege ferenda interpretations. The doctrinal methodology, although has long been the paradigm in the discipline of law, is progressively criticised as inflexible and inward-looking, in need of an interdisciplinary turn.[16] In the meantime, EU institutions more often than not examine AI-related legal issues in unison with ethical perspectives.[17] Thus, the joint approach of philosophical analysis and doctrinal legal research in this dissertation is justified, on the one hand, by the overlap of ethical and legal considerations, as this has also been observed by the EU, and, on the other hand, by the wish to ensure that the law is not read in a vacuum, but is instead embedded in a wider context.

The exposition of the EU framework is premised upon policy outputs of EU bodies, including, but not limited to, the European Parliament, the European Commission, the European Council, and the European Economic and Social Committee. Regarding soft law instruments consulted, these include both the ones officially recognised by Article 288 of the Treaty on the Functioning of the European Union (TFEU), i.e. recommendations and opinions, and unofficial ones, i.e. briefings, reports, guidelines, various communications and strategy

[16] Rob van Gestel and Hans-Wolfgang Micklitz, 'Why Methods Matter in European Legal Scholarship: Methods in European Legal Scholarship', *European Law Journal* 20, no. 3 (May 2014): 292–316, https://doi.org/10.1111/eulj.12049.

[17] Mady Delvaux, 'Report with Recommendations to the Commission on Civil Law Rules on Robotics' (Strasbourg: Committee on Legal Affairs, European Parliament, 27 January 2017), http://www.europarl.europa.eu/sides/getDoc.do?type=REPORT&reference=A8-2017-0005&language=EN; Directorate-General for Communications Networks, Content and Technology, 'Artificial Intelligence for Europe', Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the regions (Brussels: European Commission, 25 April 2018), https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237&rid=2.

declarations.[18] Concerning hard law, it is primarily instantiated in the Treaties, general principles, and the Charter of Fundamental Rights of the European Union (CFREU) and secondarily in Regulations, Directives, and Decisions (Article 288 TFEU).[19]

Poignantly, legally enforceable policies targeted to AI are not yet a reality on the EU frontier. However, by virtue of the data-driven character of AI, this dissertation substantially depends on EU legislation concerning data protection, namely the Regulation (EU) 2016/679 (General Data Protection Regulation, GDPR).[20] Its provisions are elucidated by interpretative guidelines issued by the Article 29 Data Protection Working Party (WP29), an advisory body now replaced by the European Data Protection Board.[21] Supplementary cross-references will be given to the Directive (EU) 2016/680 (Police Directive).[22] Secondary literature in the form of scholarly papers and conference proceedings is used to buttress and evaluate the relevant policies, while journalistic articles are cited to sketch the topicality of the issues examined.

It is noteworthy that, compared to the exponential growth observed in the constellation of AI, ML, and Big Data technologies during the last years, the legal framework and ancillary structures entrusted with their regulation have not evolved as rapidly. This discrepancy between the rate of technological progress and that of policy reforms is noticeable–albeit not confined–in EU jurisdiction, inasmuch as AI-focused regulation and case law is absent. This is summarised as 'the pacing problem of law'.[23]

---

[18] 'Consolidated Versions of the Treaty on European Union and the Treaty on the Functioning of the European Union', Pub. L. No. 2008/C 115/01, OJ C 115 (2008), https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1544456032916&uri=CELEX:C2008/115/01.

[19] 'Charter of Fundamental Rights of the European Union', Pub. L. No. 2007/C 303/01, OJ C 303 (2007), https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:12007P/TXT.

[20] 'Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)', Pub. L. No. 32016R0679, OJ L 119 (2016), http://data.europa.eu/eli/reg/2016/679/oj/eng.

[21] The European Data Protection Board endorsed all GDPR-related reports of the WP29: European Data Protection Board, 'Endorsement 1/2018', accessed 11 November 2018, https://edpb.europa.eu/sites/edpb/files/files/news/endorsement_of_wp29_documents.pdf.

[22] 'Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data by Competent Authorities for the Purposes of the Prevention, Investigation, Detection or Prosecution of Criminal Offences or the Execution of Criminal Penalties, and on the Free Movement of Such Data, and Repealing Council Framework Decision 2008/977/JHA', Pub. L. No. 32016L0680, OJ L 119 (2016), http://data.europa.eu/eli/dir/2016/680/oj/eng.

[23] Gary E. Marchant, Braden R. Allenby, and Joseph R. Herkert, *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem* (Dordrecht: Springer Science+Business Media B.V., 2011), http://0-dx.doi.org.fama.us.es/10.1007/978-94-007-1356-7.

Research significance

The ethical and legal implications of AI are not novel focal points in the legal discourse. Since the 1980s, scholars such as Lehman-Wilzig, Willick, and Solum have defended their views on the pitfalls of AI under US law, with no counterarguments eclipsing.[24] However, in accordance with the development of the Information and Communication Technology sector, AI faces a process of constant refinement, with new legal and ethical entanglements thereof resulting. Among such developments, the enhancement of AI with ML and Big Data has caused unprecedented difficulties in explaining how AI systems turn their inputs to conclusions, which is especially disturbing when these conclusions are discriminatory. Designated as the 'black box' problem of AI, its social extensions became evident in the introductory vignettes, as it interferes with individuals' employment prospects, sentencing decisions, freedom of movement, entertainment and communication needs.

Scholarly approaches to this issue are usually confined to a legal or ethical viewpoint. Mittelstadt et al. and Martin succinctly review ethical concerns caused by decision-making algorithms in their respective papers *The Ethics of Algorithms: Mapping the Debate* and *Ethical Implications and Accountability of Algorithms*.[25] In the legal sphere, the papers which garnered most attention regarding automated decisions in the GDPR are *European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"* by Goodman and Flaxman and its rebuttal in *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation* by Wachter et al.[26] Given the complexity of the problem at stake, though, solutions extending beyond rigid disciplinary boundaries are more likely to be sustainable. Additionally, there is yet for common ground to be found about the apposite treatment of AI bias not only among scholars but among all stakeholders involved, meaning academia, industry, governmental actors, and end users.

---

[24] Sam N. Lehman-Wilzig, 'Frankenstein Unbound', *Futures* 13, no. 6 (December 1981): 442–57, https://doi.org/10.1016/0016-3287(81)90100-2; Marshall S. Willick, 'Artificial Intelligence: Some Legal Approaches and Implications', *AI Magazine* 4, no. 2 (15 June 1983): 5–16, https://doi.org/10.1609/aimag.v4i2.392; Lawrence B. Solum, 'Legal Personhood for Artificial Intelligences', *North Carolina Law Review* 70, no. 4 (1 April 1992): 1231–87.

[25] Brent Daniel Mittelstadt et al., 'The Ethics of Algorithms: Mapping the Debate', *Big Data & Society* 3, no. 2 (December 2016): 205395171667967, https://doi.org/10.1177/2053951716679679; Kirsten Martin, 'Ethical Implications and Accountability of Algorithms', *Journal of Business Ethics*, 7 June 2018, https://doi.org/10.1007/s10551-018-3921-3.

[26] Bryce Goodman and Seth Flaxman, 'European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"', *AI Magazine* 38, no. 3 (2 October 2017): 50–57, https://doi.org/10.1609/aimag.v38i3.2741; Sandra Wachter, Brent Mittelstadt, and Luciano Floridi, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation', *International Data Privacy Law* 7, no. 2 (3 June 2017): 76–99, https://doi.org/10.1093/idpl/ipx005.

These disputes also reflect divergent ethical and legal traditions. So far, the burgeoning discourse on AI was US-focused, with the articles *Big Data's Disparate Impact* by Barocas and Selbst and *Technological Due Process* by Citron leading the discussion.[27] Nonetheless, as AI technologies are notoriously data-hungry and thereby raise a host of issues on data protection, the recent advent of the GDPR, characterised as a 'Copernican revolution', has brought EU policies to light.[28] Its influence extends beyond the EU and is discernible even in the California Consumer Privacy Act in the US.[29] Therefore, there is pivotal comparative interest to a presentation of the relevant EU provisions.

Of course, this analysis does not intend to be exhaustive or decisive. Rather, it aims at alleviating the 'pacing problem' between law and technological progress. As Marchant puts it, there are two routes to addressing the pacing problem: halt the appearance and advancement of evolving technologies or embrace novel mechanisms and approaches to accommodate their regulation.[30] Given, firstly, that the former option entails the side effect of blocking positive aspects of technological products and, secondly, that uncritically and cursorily adopting the latter option would be short-sighted, it suffices for this dissertation to provide an ethical and legal roadmap that will bring the literature even slightly closer to the thriving AI field.

[27] Solon Barocas and Andrew D. Selbst, 'Big Data's Disparate Impact', *California Law Review* 104 (2016): 671–732, https://doi.org/10.15779/z38bg31; Danielle Keats Citron, 'Technological Due Process', *Washington University Law Review* 85, no. 6 (2008): 1249–1313.
[28] Christopher Kuner, 'The European Commission's Proposed Data Protection Regulation: A Copernican Revolution in European Data Protection Law', Bloomberg BNA Privacy and Security Law Report (Rochester, NY, 6 February 2012), https://papers.ssrn.com/abstract=2162781.
[29] Reece Hirsch et al., 'California's New, GDPR-Like Privacy Law Is A Game-Changer', Bloomberg Law, 11 July 2018, https://news.bloomberglaw.com/privacy-and-data-security/insight-californias-new-gdpr-like-privacy-law-is-a-game-changer.
[30] Marchant, Allenby, and Herkert, *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem*, 19–20.

Chapter 1: Technical overview

1.1 Definitions of AI

References to AI bring to mind narratives such as the Frankenstein's monster conceived by Mary Shelley, the Ovidian myth of Pygmalion in *Metamorphoses*, or the Jewish legend of the Golem of Prague. More often, the word 'robot' is used alternatively, yet mistakenly as will be later illuminated, taken from Karel Čapek's 1920 play *R.U.R.* or *Rossum's Universal Robots*. This is how the European Parliament begins its Resolution *Civil Law Rules on Robotics* to illustrate the persistence of AI-related issues among human concerns.[31] Indeed, as far back as the conception of the *Iliad,* Homer describes how Hephaestus crafted golden gynoids to be his assistants (18.415), whereas in *Argonautica* we read that he manufactured Talos, a mechanical colossus patrolling the shores of Crete (IV, 11.1638).[32] In modern terms, Talos would be the first 'killer AI robot'.[33] However, the public should be careful not to lend credibility to these depictions. Such portrayals, spurred by sensationalist media coverage or artistic imagination, are as distant from reality as possible and should be counteracted by realistic descriptions of AI, if one wants to rationally deliberate its ethical and legal bifurcation.

EU institutions embrace the goal of realistically describing AI, but agree in that no single accepted definition has yet been formulated.[34] Thus, they embark on constructing their own. Their attempted definitions share a triad of necessary conditions to characterise a technological output as AI.[35]

---

[31] European Parliament et al., 'European Parliament Resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))', *Official Journal of the European Union*, C, 61, no. 252 (18 July 2018): 239–257.

[32] 'Homer, Iliad, Book 18', accessed 11 October 2018, http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.01.0134%3Abook%3D18; 'The Argonautica, by Apollonius Rhodius', accessed 11 October 2018, http://www.gutenberg.org/files/830/830-h/830-h.htm.

[33] Stephen Cave and Kanta Dihal, 'Ancient Dreams of Intelligent Machines: 3,000 Years of Robots', *Nature* 559 (25 July 2018): 473, https://doi.org/10.1038/d41586-018-05773-y.

[34] Peter Bentley et al., *Should We Fear Artificial Intelligence?: In-Depth Analysis* (Brussels: Scientific Foresight Unit, European Parliament, 2018), http://www.europarl.europa.eu/RegData/etudes/IDAN/2018/614547/EPRS_IDA(2018)614547_EN.pdf; Catelijne Muller, 'Artificial Intelligence – The Consequences of Artificial Intelligence on the (Digital) Single Market, Production, Consumption, Employment and Society', Opinion (Brussels: European Economic and Social Committee, 31 May 2017), https://www.eesc.europa.eu/en/our-work/opinions-information-reports/opinions/artificial-intelligence.

[35] Vincent Reillon, 'Understanding Artificial Intelligence', Briefing (European Parliamentary Research Service, European Parliament, January 2018), http://www.iberglobal.com/files/2018/Understanding_AI.pdf; Directorate-General for Communications Networks, Content and Technology, 'Artificial Intelligence for Europe'.

Firstly, AI consists of sets of algorithms which analyse data and statistical relations. The term algorithm, deriving from the 12th-century Persian scholar al-Khwārizmī (Latinised as Algoritmi) and denoting the step-by-step performance of elementary arithmetic, has come to mean a series of defined steps undertaken to produce particular outputs.[36] In ethical and legal discussions, though, algorithms are not mentioned solely as abstract mathematical constructs. Today, algorithms come with their implementation in a computer program or information system.[37] Of particular interest in this dissertation is a specific class of algorithms, namely those that take decisions, such as what is the best action to take in certain situations or the best interpretation of the provided data, to support or replace human decision-making.[38] Diakopoulos picks out the following types of algorithmic decisions: prioritisations accentuate particular information items over others; classifications assort entities into classes based on their characteristics; associations limn connections among entities; filtering encloses or discloses information in conformity with benchmarks.[39]

Secondly, AI has the ability to perform a task. EU definitions specify that the task must be goal-oriented, without any restrictions on what constitutes a possible goal for the AI.[40] The European Commission and the European Parliament in their definitions further include some degree of autonomy in task performance.[41] Conversely, the European Group on Ethics in Science and New Technologies (EGE) disapproves of this requirement, as falsely associating the absence of human supervision with autonomy. According to EGE, autonomy implies adhering to goals in a deliberate, independent way, which results solely from self-awareness; an awareness that, contrary to humans, AI is currently thereof deprived.[42]

Thirdly, tasks performed by AI must be such that would otherwise require human intelligence. Tasks commonly attributed to intelligence are *'reasoning, the gathering of information, planning, learning, communicating, manipulating, detecting and even creating,*

[36] Rob Kitchin, 'Thinking Critically about and Researching Algorithms', *Information, Communication & Society* 20, no. 1 (2 January 2017): 14–29, https://doi.org/10.1080/1369118X.2016.1154087.
[37] Mittelstadt et al., 'The Ethics of Algorithms'.
[38] Mittelstadt et al.
[39] Nicholas Diakopoulos, 'Algorithmic Accountability: Journalistic Investigation of Computational Power Structures', *Digital Journalism* 3, no. 3 (4 May 2015): 398–415, https://doi.org/10.1080/21670811.2014.976411.
[40] Bentley et al., *Should We Fear Artificial Intelligence?*
[41] Directorate-General for Communications Networks, Content and Technology, 'Artificial Intelligence for Europe'; European Parliament et al., 'Civil Law Rules on Robotics'.
[42] European Group on Ethics in Science and New Technologies, 'Artificial Intelligence, Robotics and "Autonomous" Systems', Statement (Luxembourg: Directorate-General for Research and Innovation, European Commission, 30 April 2018), https://publications.europa.eu/en/publication-detail/-/publication/dfebe62e-4ce9-11e8-be1d-01aa75ed71a1/language-en.

dreaming and perceiving'.[43] Although these cognitive functions need not manifest all at once in an AI system, they must do so to an extent at least comparable to human intelligence. Concurrently, a condition adopted by the Commission is the ability of AI to learn from its environment, and thereby adapt to it and improve its performance.[44] Overall, this combination of technological abilities with properties related to human intelligence is the main feature distinguishing AI from other technologies, while being the feature most inconsistently described in EU communications.

Taken together, these three core elements give the working definition of this dissertation:

> **AI systems are sets of algorithms that analyse data and statistical relations to perform a goal-oriented task which would otherwise require human intelligence.**

In order for these systems to be built, one or more scientific fields are employed under the blanket term AI development: through cognitive computing, computer scientists develop algorithms with the ability to reason; through Machine Learning, they create algorithms that can teach themselves tasks; within the contours of augmented intelligence, they explore human-machine cooperation; and in AI robotics, they produce AI embedded in robots.[45] Concerning the last field, EU policy-makers are proactive in clarifying a common confusion between robotics and AI.[46] AI systems with the aforementioned core features are embedded either in purely software systems, in which case they operate in the virtual sphere as software robots/(soft)bots (e.g. virtual assistants, image or speech analysis software, search engines, chatbots) or in hardware systems, in which case they are embodied in a physical structure (e.g. robots, self-driving cars, drones, Internet of Things).[47] Robots do not necessarily function based on AI. However, as more and more AI systems perform their manual and cognitive tasks through the physical embodiment of robots, their interplay gets tighter, giving rise to the field of AI robotics. Although the European Parliament in its Resolution makes the unfortunate decision to address only embodied AI, thus leaving outside its scope a wide range of AI systems, all other EU policies address embodied and non-embodied AI in tandem.[48]

---

[43] Bentley et al., *Should We Fear Artificial Intelligence?*
[44] Directorate-General for Communications Networks, Content and Technology, 'Artificial Intelligence for Europe'.
[45] Muller, 'Artificial Intelligence – The Consequences of Artificial Intelligence on the (Digital) Single Market, Production, Consumption, Employment and Society'.
[46] Delvaux, 'Report with Recommendations to the Commission on Civil Law Rules on Robotics'.
[47] Directorate-General for Communications Networks, Content and Technology, 'Artificial Intelligence for Europe'.
[48] European Parliament et al., 'Civil Law Rules on Robotics'; Directorate-General for Communications Networks, Content and Technology, 'Artificial Intelligence for Europe'.

Normative efforts vis-à-vis such artificial agents may be recent in law, but not fiction. In the General Principles of its Resolution, the European Parliament appeals to Asimov's Laws, which should govern the conduct of developers and users of autonomous, self-learning robots, like the ones embodying AI.[49] In Isaac Asimov's collection of stories, *I, Robot*, we read the following quotes:

> *'One, a robot may not injure a human being, or, through inaction, allow a human being to come to harm.'*
>
> *'Two, […] a robot must obey the orders given it by human beings except where such orders would conflict with the First Law.'*
>
> *'And three, a robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.'*[50]

Although fictional, these three rules got traction in the policy-making territory, as witnessed by the European Parliament's decision to include them in the Resolution and their frequent reference in relevant scholarship.

1.2 The Turing Test and the Chinese Room

As seen in part 1.1, the third necessary condition for the definition of AI is the performance of tasks which would otherwise require human intelligence. To shed light on this requirement, it is useful to explain the Turing Test, as this is tersely mentioned in EU reports, but also its counter thought experiment, the Chinese Room, which is missing in the reports.[51]

In 1950, British mathematician Alan Turing published *Computing machinery and intelligence*.[52] In this influential paper, Turing firstly asked whether machines can think. Considering this research question obscure, he replaced it with a test on whether a machine could act in ways which would persuade a human that it could think.[53] If the test was positive, meaning that a human interrogator was convinced of communicating with another human instead of a machine during 30% of their conversation, then the machine would have achieved true intelligence.[54] The main premise of this 'Turing Test' or 'Imitation Game' is that if a machine acts like an intelligent human being, then it must truly be intelligent.

---

[49] European Parliament et al., 'Civil Law Rules on Robotics'.
[50] Isaac Asimov, *I, Robot*, accessed 11 October 2018, https://www.ttu.ee/public/m/mart-murdvee/Techno-Psy/Isaac_Asimov_-_I_Robot.pdf.
[51] Reillon, 'Understanding Artificial Intelligence'.
[52] A. M. Turing, 'Computing Machinery and Intelligence', *Mind* LIX, no. 236 (1950): 433–60, https://doi.org/10.1093/mind/LIX.236.433.
[53] Turing.
[54] Stuart J. Russell, Peter Norvig, and Ernest Davis, *Artificial Intelligence: A Modern Approach*, 3rd ed, Prentice Hall Series in Artificial Intelligence (Upper Saddle River: Prentice Hall, 2010), 1021.

Although the Turing Test is still relevant and takes place under the Loebner Prize competition, no AI has ever managed to pass it.[55] The successful AI system should display a range of technically demanding and specialised features, so it is reasonable for AI researchers not to concentrate their efforts on building machines that pass the Turing Test. By recollecting technological inventions that shaped our society, e.g. 'artificial flight', it becomes evident that their creators were more preoccupied with delving into and applying engineering theories than building *'machines that fly so exactly like pigeons that they can fool even other pigeons'*.[56]

The most important counterargument to the Turing Test has been the 'Chinese Room'. The American philosopher John Searle explicated this thought experiment in *Minds, brains, and programs* in 1980 and reformulated it in his 1990 paper *Is the Brain's Mind a Computer Program?*.[57] Searle imagines himself, an English speaker with no fluency in Chinese, locked in a room. He is given one batch of Chinese symbols and one batch of rules written in English that show how certain Chinese symbols correspond to other Chinese symbols. The rules do not refer to the meaning of the symbols, but only identify them by their shapes, as in 'the symbol with such and such lines goes with the symbol of this or that shape'. From outside the room, Chinese speakers pass him batches of Chinese symbols. By applying the aforementioned rules, he gives them back batches of other Chinese symbols.

In this speculative scenario, the rules written in English symbolise the computer program and Searle is the computer.[58] The first batch of Chinese symbols given to the computer is the database, the second batch coming from the outside speakers are questions, and the third batch which the computer returns are answers.[59] By following the program, the computer gives correct responses in Chinese and exhibits behaviour which to an external observer is indistinguishable from that of a real Chinese speaker. In this way, the computer satisfies the Turing Test. Nevertheless, Searle or the computer does not understand Chinese at all, as in spite of the correct manipulation of symbols, no meaning is attached to them.[60] In other words, the computer does not grasp the semantics of the symbols, without which we cannot corroborate that something or someone is actually thinking.

---

[55] Luciano Floridi, Mariarosaria Taddeo, and Matteo Turilli, 'Turing's Imitation Game: Still an Impossible Challenge for All Machines and Some Judges—An Evaluation of the 2008 Loebner Contest', *Minds and Machines* 19, no. 1 (February 2009): 145–50, https://doi.org/10.1007/s11023-008-9130-6.
[56] Russell, Norvig, and Davis, *Artificial Intelligence*, 3.
[57] John R. Searle, 'Minds, Brains, and Programs', *Behavioral and Brain Sciences* 3, no. 03 (September 1980): 417, https://doi.org/10.1017/S0140525X00005756; John R. Searle, 'Is the Brain's Mind a Computer Program?', *Scientific American* 262, no. 1 (January 1990): 26–31, https://doi.org/10.1038/scientificamerican0190-26.
[58] Searle, 'Is the Brain's Mind a Computer Program?'
[59] Searle.
[60] Searle.

With this thought experiment, Searle refutes the thesis of Strong AI, which implies that AI systems could have conscious mental states because of their ability to manipulate formal symbols, and confines it to the Weak AI theory, according to which AI can be programmed only to outwardly exhibit intelligent behaviour.[61]

1.3 Types of AI

Corresponding to Searle's distinction between Weak and Strong AI, the European Economic and Social Committee (EESC) and the European Parliamentary Research Service (EPRS) differentiate between Narrow and General AI.[62]

Narrow AI, nicknamed as 'tool AI', denotes machines able to carry out specific tasks or sets thereof.[63] Although Narrow AI can surpass the respective human performance in these tasks, it differs from human intelligence, as the latter allows for transferring skills from one task to another. For instance, Narrow AI is programmed to recognise a specific kind of animal or play a specific kind of game. The combination of such one-task AI systems results in composite technological outputs, such as self-driving cars.[64]

On the contrary, General AI is endowed with broad cognitive abilities, even consciousness. As it performs any mental task attributed to human beings, its intelligence would be tantamount to that of a human.[65] Although General AI, referred to as 'real AI', was the initial impetus of AI research, it is admittedly a tall order for developers. One of the most internationally notable endeavours in the field, the 10-year-long Human Brain Project, is supported by the Commission's Future & Emerging Technologies Programme, which offers large-scale investments in transformative research.[66]

---

[61] Searle.
[62] Reillon, 'Understanding Artificial Intelligence'; Muller, 'Artificial Intelligence – The Consequences of Artificial Intelligence on the (Digital) Single Market, Production, Consumption, Employment and Society'.
[63] Muller, 'Artificial Intelligence – The Consequences of Artificial Intelligence on the (Digital) Single Market, Production, Consumption, Employment and Society'.
[64] Reillon, 'Understanding Artificial Intelligence'.
[65] Reillon.
[66] Seth D. Baum, 'A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy', *Global Catastrophic Risk Institute Working Paper 17-1*, 2017, https://doi.org/10.2139/ssrn.3070741; 'Human Brain Project Flagship', Digital Single Market, accessed 4 November 2018, https://ec.europa.eu/digital-single-market/en/human-brain-project.

1.4 The development of AI

Cast broadly, the historical development of AI is divided into three stages reflecting the prevalence of three techniques: symbolic AI, Machine Learning, and Deep Learning.

1.4.1 Symbolic AI

The publication of Turing's seminal paper *Computing machinery and intelligence* in 1950 designates the first stage in AI development.[67] Following this, in 1956 John McCarthy and Marvin Minsky hosted the historic Dartmouth Summer Research Project on Artificial Intelligence, convening the figures that would dominate AI in the years to come.[68] Quoting McCarthy and the phrase in which the term Artificial Intelligence was first coined:

> *'We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.'*[69]

Subsequently, the 1960s are characterised by the development of symbolic AI in the US and the UK.[70] Based on the theory beneath symbolic AI, intelligent human behaviour is dissected into a set of smaller problems. The cognitive steps that a human expert would take to solve these problems are identified and translated into successive rules of logic.[71] This happens through the cooperation of two systems. A knowledge base, which translates facts about the world to logical symbols, is used by an inference engine, which applies logical rules to deduce new knowledge from it in the form of algorithms.[72] By inserting these algorithms into a computer, it imitates the intelligent behaviour which the algorithms symbolically described. Owing to the algorithmic simulation of a human expert's decision-making sequence, this is called an expert system.[73]

In practice, describing in logical terms all the rules and possible ways to solve a problem proved arduous and dependent on more computational power than what was then available.

---

[67] Reillon, 'Understanding Artificial Intelligence'.
[68] Russell, Norvig, and Davis, *Artificial Intelligence*, 17.
[69] Russell, Norvig, and Davis, 17.
[70] Reillon, 'Understanding Artificial Intelligence'.
[71] Reillon.
[72] Reillon.
[73] Reillon.

Insufficient computational power meant that computers of the time were lacking the necessary storage and information processing capacities. Despite their overconfident predictions, AI researchers failed to meet their announced goals. Consequently, the initial enthusiasm for AI abated and, with that, the respective governmental funding in the UK and the US, leading to the so-called 'AI winters' at the beginning of the 1970s and the end of the 1980s.[74]

1.4.2 Machine Learning

In 1997, IBMs' AI Deep Blue defeated the chess grandmaster Garry Kasparov, galvanising a renewed interest in AI, which manifested as the second stage of AI development.[75] During the 2000s, computer scientists relied on a distinctive trait of human intelligence, the ability to learn, conceived as the ability to utilise experience to improve behaviour.[76] Based on this conception, they established the methodology of Machine Learning (ML), which enables algorithms to learn from their experience and ameliorate. ML techniques achieve this by using input data to identify patterns and create their own models for predicting future outputs. Depending on the domain for which they are designed, their accuracy in classification, and the availability of computational resources, ML models take the form of decision trees, logistic regression, Naïve Bayes, neural networks, or combinations of these in 'model ensembles'.[77]

Specifically, neural network models simulate the operation of a human brain and are based upon the creation of artificial neurons.[78] When these artificial neurons are intertwined in multiple layers, they form an artificial neural network. Each time this network receives an input signal, it processes it to produce an output signal.[79] What is interesting in ML is that the machine autonomously modifies the interactions in its network so that it consistently gives the expected output every time it receives an input.[80] This constitutes its process of learning or training. For instance, if the goal is to train AI to recognise images of cats, one needs to provide it with a large quantity of training data, in this case different images. Upon receiving these images as input, the network produces either the output that this is an image of a cat or the output that this is not an image of a cat. If its output is correct, it will strengthen its network

[74] Reillon.
[75] Reillon.
[76] Reillon.
[77] Jenna Burrell, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms', *Big Data & Society* 3, no. 1 (5 January 2016): 1–12, https://doi.org/10.1177/2053951715622512.
[78] Reillon, 'Understanding Artificial Intelligence'.
[79] Reillon.
[80] Reillon.

interactions. If it is incorrect, it will readjust its network interactions based on the new information. Through multiple repetitions, the machine is able to provide the correct output.[81]

ML is divided into supervised and unsupervised, although in practice the distinction is not stringent. In the first case, a developer labels the input data and their corresponding outputs.[82] In the second case, input data are unlabelled and the machine finds associations among them on its own.[83] Especially in unsupervised ML, due to the absence of human direction, it is impossible to identify which features of the input data were used by the ML model to reach its final output and in what ways. Consequently, predictions or explanations of how an input is or will be handled by ML models become unattainable.[84] This 'black box' method, as we shall see, raises a plethora of ethical and legal challenges. On the other side, unsupervised ML is closer to the model of human intelligence, which increases its possibilities for long-term applicability. As LeCun et al. point out, *'we discover the structure of the world by observing it, not by being told the name of every object'*.[85]

1.4.3 Deep Learning and Big Data

From the mid-2000s until nowadays, the third stage of AI development is characterised by rapid progress in ML and its advancement to Deep Learning (DL). Deep Learning builds on the previous ML model and, by multiplying the layers of artificial neurons and combining different ML techniques, is able to solve more complex problems.[86] The distinctive feature of DL is that these layers are not designed by developers, but created by the system itself through its learning process.[87]

Given the increased production of data in the digital sphere nowadays, ML and DL systems access a wider pool of available training data which help them improve. This is especially facilitated by Big Data, meaning data and processing with the following attributes, known as the 'four Vs of Big Data': high volume of data; high velocity in their processing; veracity (or uncertainty) of data; and variety in their types.[88] Although Big Data are a valuable

[81] Reillon.
[82] Reillon.
[83] Reillon.
[84] Mittelstadt et al., 'The Ethics of Algorithms'.
[85] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, 'Deep Learning', *Nature* 521, no. 7553 (May 2015): 436–44, https://doi.org/10.1038/nature14539.
[86] Reillon, 'Understanding Artificial Intelligence'.
[87] LeCun, Bengio, and Hinton, 'Deep Learning'.
[88] Brent Daniel Mittelstadt and Luciano Floridi, 'The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts', *Science and Engineering Ethics* 22, no. 2 (April 2016): 303–41, https://doi.org/10.1007/s11948-015-9652-2.

asset, traditional data analysis methods have been insufficient in handling them. This is where AI, in the form of ML and DL, is key to unlocking the insights that Big Data harbour. On top of that, the growing availability of computer processing power alongside the rise of cloud computing and the increased sophistication of algorithms contribute to the current exponential growth of AI. In computer science, Moore's Law is commonly invoked, stating that *'the number of transistors in a dense integrated circuit doubles approximately every two years'*.[89] Roughly, this implies that over time the hardware required to run the same technology is halved; thus, computational power and technological progress in toto rapidly escalate. AI development has not only followed this rate but is slated to soon outpace it.[90]

Despite the meteoric rise of AI, this historical overview should not be seen as merely informative, but as conducive to deriving conclusions applicable to present and future AI development. Witnessing the non-linear trajectory of AI, with alterations between excitement and sceptic 'AI winters', is important to evaluate it from a realistic standpoint. As evidently happened in the past, it is not impossible for the hype of AI to give its place to a suspension of relevant activities. Moreover, in the course of these three stages, AI development is moving from a primarily governmental and academic endeavour to the private sector, which implies an increase in investments but a decrease in accountability. Finally, although the US and the UK led AI development in its inception, the accelerating Chinese AI sector is deemed threatening to the other countries and likely to overshadow them in the following decades.[91]

1.5 Current state-of-the-art

EU reports acknowledge the ubiquitous uptake of AI systems in all aspects of life, both for public and private usage. Their applications manifest as optimised functions of daily products, such as identifying spam emails, providing customer support in the form of virtual assistants or purchasing suggestions, and translating texts.[92] Equally, they span disruptive and life-

[89] 'Processing Power beyond Moore's Law | News', CORDIS | European Commission, accessed 4 November 2018, https://cordis.europa.eu/news/rcn/129281_en.html.
[90] 'Artificial Intelligence Is Awakening the Chip Industry's Animal Spirits', *The Economist*, 7 June 2018, https://www.economist.com/business/2018/06/07/artificial-intelligence-is-awakening-the-chip-industrys-animal-spirits.
[91] Laurent Probst et al., 'USA-China-EU Plans for AI: Where Do We Stand?' (Digital Transformation Monitor, European Commission, January 2018), https://ec.europa.eu/growth/tools-databases/dem/monitor/sites/default/files/DTM_AI%20USA-China-EU%20plans%20for%20AI%20v5.pdf; Directorate-General for Communications Networks, Content and Technology, 'Artificial Intelligence for Europe'.
[92] Directorate-General for Communications Networks, Content and Technology, 'Artificial Intelligence for Europe'; Delvaux, 'Report with Recommendations to the Commission on Civil Law Rules on Robotics'.

determining interventions, such as self-driving vehicles, clinical diagnoses, lethal autonomous weapons, and deep-sea or space exploration robots.[93] To monitor Member States' readiness regarding AI, the Commission published a workshop report on *The European AI Landscape* and launched AI Watch, a knowledge service that will steadily conduct such monitoring.[94]

Despite its successful application in myriad specialised tasks, AI cannot exhibit the full range of human mental states. AI systems are still unable to display common sense, have affective feelings, or share their human users' intentions. This means that the progress of AI has been restricted in its Narrow type, whereas only fractional steps have been achieved in General AI, which would bestow machines with commonsensical reasoning, emotions, and consciousness. Even more distant seems the scenario of a technological 'singularity', a speculative state in which a 'superintelligence' would by all measures surpass and extinguish the human species. Nonetheless, the European Parliament's Resolution, quite hastily, purports in Recital P that it is possible for AI to surpass human intellectual capacity in the long-term.[95] Although it is not logically impossible for such a development to occur, the EESC is more focused in its policy suggestions by warning that it is urgent to examine the real-life, presently occurring implications of Narrow AI instead of the fictional ones of General AI.[96]

1.6 Conclusion

Although the definitions of AI offered in EU policy reports share the three core conditions of algorithmic structure, goal-oriented task performance, and alternatively required human intelligence, there is still a long way to go before all EU bodies, or even better Member States, end up with a stable definition. Establishing such definition is no easy feat, though. It has to balance between accuracy, so as to precisely define the scope of relevant laws, and flexibility, so as not to stifle innovation. As far as this dissertation is concerned, having explained the types and stages of AI, the focus will be on Narrow AI in the form of ML, powered by Big Data. Moving on to the next Chapter, technical traits examined here reveal their linkages to ethical issues.

---

[93] Directorate-General for Communications Networks, Content and Technology, 'Artificial Intelligence for Europe'; Delvaux, 'Report with Recommendations to the Commission on Civil Law Rules on Robotics'.
[94] Charlotte Stix, 'The European AI Landscape' (Brussels: Directorate-General for Communications Networks, Content and Technology, European Commission, 18 April 2018), https://ec.europa.eu/digital-single-market/en/news/european-artificial-intelligence-landscape; 'AI Watch', Knowledge for policy, 28 November 2018, https://ec.europa.eu/knowledge4policy/ai-watch_en.
[95] European Parliament et al., 'Civil Law Rules on Robotics'.
[96] Muller, 'Artificial Intelligence – The Consequences of Artificial Intelligence on the (Digital) Single Market, Production, Consumption, Employment and Society'.

Chapter 2: The ethical challenges of AI

2.1 Introduction

Chapter 1 delineated how AI processes data through ML techniques. The consequential nature of this processing, meaning that it results in evidence for decision-making or decisions per se and in accordingly motivating actions, raises ethical concerns. Although technological artefacts used to be ethically problematic when they malfunctioned, this Chapter demonstrates that the ethical neutrality of AI is not guaranteed even when it works as designed. Following Mittelstadt et al. in their division of ethical concerns into epistemic and normative ones, parts 2.2, 2.3, 2.4, 2.5, 2.7, and 2.8 reflect epistemic concerns, while part 2.6 discusses normative ones.[97] In all parts, ethical concerns relate to technical features of the AI, ML, and Big Data convergence in decision-making contexts.

This Chapter fixates on the two-pronged notion of 'bias'. In its neutral sense, bias refers simply to a preference or inclination towards a particular object, subject, or area, as in 'x's writing is biased towards theological doctrines'. ML models are by design biased in this neutral sense, because, in order to generate outputs for input data beyond the training sample, they form generalising assumptions, called 'inductive biases'.[98] Without such biases, ML models could not produce predictions for cases they have not encountered before. As this is an inherent trait of ML, we cannot expect it to be unbiased in this sense.[99]

In the second, moralised use of the term, bias denotes an '*[i]nclination or prejudice for or against one person or group, especially in a way considered to be unfair*', as in 'y is biased against the Jews'.[100] What underpins this kind of bias is forming beliefs or acting towards persons in a differential way because of their perceived characteristics, such as gender, race, or sexual orientation. Such biases manifest in positive (bias for) or negative (bias against) forms, wherein the former leads to favouritism and the latter to unfair discrimination. Hence, biases in this sense are morally problematic and, especially when systematically exhibited, precarious to democratic communities. In what follows, it is argued that the AI, ML, and Big Data

[97] Mittelstadt et al., 'The Ethics of Algorithms'.
[98] Thomas G. Dietterich and Eun Bae Kong, 'Machine Learning Bias, Statistical Bias, and Statistical Variance of Decision Tree Algorithms' (Oregon State University, 1995), http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.38.2702&rep=rep1&type=pdf.
[99] Dietterich and Kong.
[100] 'Definition of Bias', Oxford Dictionaries | English, accessed 7 November 2018, https://en.oxforddictionaries.com/definition/bias.

assemblage is biased not only in the first, neutral sense but also in the second, moralised one, which is also the one colloquially used.

At first glance, it seems odd to discuss ethics in respect of AI, as morality typically presupposes the existence of a human agent.[101] Notwithstanding the flourishing literature on Artificial Moral Agents, referring to the possibility of creating General AI with moral consciousness, this dissertation is framed around Narrow AI, which exhibits autonomy only in a limited technological, not moral sense.[102] However, even without moral consciousness, AI algorithms are 'actants', parts of a system consisting of material and non-material/human actors aimed at the attainment of a goal.[103] According to Latour, moral responsibility lies neither on the end of the human actor nor on the end of the technological system itself.[104] Rather, the human actor has the flexibility to delegate tasks to the system in varying degrees, without the overall responsibility being diminished or extinct.[105] This is why, when examining the ethical or not operation of an AI system, one should speculate the morality of the equivalent human action, had this delegation never happened. This is even more applicable to AI systems which undertake socially significant functions by replacing human judgements on housing, justice, education, and employment matters.

2.2 The promise of objectivity

Promulgating that the biases of AI beget epistemic and normative challenges counteracts the conventional wisdom that it is preferred over humans in decision-making contexts because of its ethical neutrality.

AI systems are heralded as liberating decision-making from human errors and prejudices. Void of the emotional dispositions inherent in human attitudes, they are purportedly established upon solid and rational statistical bases. Specifically, the process of employing ML algorithms to parse Big Data and thereby arrive at decisions has striking empiricist undertones. Back in Ancient Greece and contra Plato, who advocated that the Forms are grasped deductively, Aristotle endorsed empirical, observation-based reasoning as the means to attain

[101] Following Sidgwick, in this dissertation ethics, morality, and their derivatives are used interchangeably: Henry Sidgwick, *Outlines of the History of Ethics for English Readers* (London: Macmillan, 1886), 11, http://archive.org/details/outlinesofhistor00sidguoft.
[102] For the relevant discussion see: Wendell Wallach and Colin Allen, *Moral Machines* (Oxford University Press, 2009), https://doi.org/10.1093/acprof:oso/9780195374049.001.0001; Ryan Tonkens, 'Out of Character: On the Creation of Virtuous Machines', *Ethics and Information Technology* 14, no. 2 (June 2012): 137–49, https://doi.org/10.1007/s10676-012-9290-1.
[103] Martin, 'Ethical Implications and Accountability of Algorithms'.
[104] Martin.
[105] Martin.

knowledge of the Forms.[106] Variances to Aristotle's method were defended by Francis Bacon in *Novum Organum* and Isaac Newton in *Principia.* Bacon warned that in proceeding through deductive reasoning, people are prone to altering their reasoning process so that it fits the hypothesis under examination.[107] The sole way to evade this is induction, the use of particular empirical data to infer general conclusions for other similar cases from a bottom-up direction. Newton adhered to Bacon's ideas and rejected the use of hypotheses in scientific enquiry, as evident in his characteristic quote *'hypotheses non fingo'*.[108] Later on, the inductive method was criticised by David Hume and Karl Popper. In dialogue with Hume's theories, Popper articulated the logical and psychological problems of induction. The logical aspect posits that humans are not justified in arguing about universal truths on the grounds of particular instances that they have experienced. According to the psychological aspect, they hold such unjustified beliefs because of habit.[109]

Modern science is not devoted to either deductive or inductive reasoning; rather, it avails itself of both as auxiliaries. Be that as it may, the confluence of AI, ML, and Big Data revitalises the quandary on which should be considered the right method of enquiry. In a rhetoric by and large commercially motivated and reminiscent of empiricist views, merging these technologies is credited with the ability to transcend human deficiencies and catalyse the discovery of objective truth.[110] Especially Big Data are touted as enabling numbers to 'speak for themselves' and yielding substantial insights on their own, dispensing with the need of in advance hypotheses.[111] By force of a 'digital serendipity', they answer questions that people did not even know they had.[112] Based on the premise that the amount of data available and the quality of answers derived from these are proportionate, Big Data analysed through ML are ostensibly able to exhaustively capture reality. In addition, by virtue of their heterogeneity,

[106] Hanne Andersen and Brian Hepburn, 'Scientific Method', in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta, Summer 2016 (Metaphysics Research Lab, Stanford University, 2016), https://plato.stanford.edu/archives/sum2016/entries/scientific-method/.
[107] David J. Glass and Ned Hall, 'A Brief History of the Hypothesis', *Cell* 134, no. 3 (August 2008): 378–81, https://doi.org/10.1016/j.cell.2008.07.033.
[108] Translated as 'I frame no hypotheses'. Glass and Hall.
[109] Marisa Vasconcelos, Carlos Cardonha, and Bernardo Gonçalves, 'Modeling Epistemological Principles for Bias Mitigation in AI Systems: An Illustration in Hiring Decisions', *ArXiv:1711.07111 [Cs]*, 19 November 2017, https://doi.org/10.1145/3278721.3278751.
[110] Rob Kitchin, 'Big Data, New Epistemologies and Paradigm Shifts', *Big Data & Society* 1, no. 1 (10 July 2014): 1–12, https://doi.org/10.1177/2053951714528481.
[111] Chris Anderson, 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete', *Wired*, 23 June 2008, https://www.wired.com/2008/06/pb-theory/.
[112] Kitchin, 'Big Data, New Epistemologies and Paradigm Shifts'.

they neutralise any errors or biases that have infiltrated the data-set.[113] Thereupon, AI decision-making systems are viable remedies for the repeatedly biased and erroneous human judgements.[114]

Working together, AI, ML, and Big Data enact a modus operandi which, for its proponents, should become hegemonic, owing to its self-sufficiency and uncontested objectivity. This quest to establish legitimacy in the area of knowledge production and decision-making is neither new-fangled nor negligible. As Bourdieu has demonstrated, demarcating what counts as legitimate knowledge or not is an embodiment of power, and by delegating this demarcation to AI developers and firms they get endowed with the ability to wield such power.[115] Meanwhile, the veneer of objectivity surrounding Big Data and algorithmic decisions makes it intractable for individuals to doubt them, shaping both their perception of the world and their self-perception, due to what Delacroix calls 'anchoring effects'.[116] With a thereby distorted sense of self, humans' ethical compasses are inextricably influenced.[117]

Over and above these idealised attributes, oftentimes in the relevant discourse AI, ML, and Big Data appear to be working 'like magic', suggesting both their chimerical superpowers and their inaccessibility to the lay mind.[118] Due to increasing references of this type, boyd and Crawford acknowledged as integral in Big Data an element of mythology, which signifies:

> *'the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy.'*[119]

Brushing aside their mythological dimensions, though, AI, ML, and Big Data are socio-technical phenomena, whose development and use cannot be decoupled from a social context. Latour and Woolgar, through their two-year-long quasi-anthropological observation of interactions occurring in a laboratory, have long attested the social construction of scientific

---

[113] danah boyd and Kate Crawford, 'Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon', *Information, Communication & Society* 15, no. 5 (June 2012): 662–79, https://doi.org/10.1080/1369118X.2012.678878.
[114] Tal Zarsky, 'The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making', *Science, Technology, & Human Values* 41, no. 1 (January 2016): 118–32, https://doi.org/10.1177/0162243915605575.
[115] Diana E. Forsythe, 'Engineering Knowledge: The Construction of Knowledge in Artificial Intelligence', *Social Studies of Science* 23, no. 3 (1 August 1993): 445–77, https://doi.org/10.1177/030631279302300302.
[116] Sylvie Delacroix, 'Pervasive Data Profiling, Moral Equality and Civic Responsibility', *SSRN Electronic Journal*, 2017, https://doi.org/10.2139/ssrn.3022188.
[117] Delacroix.
[118] M. C. Elish and danah boyd, 'Situating Methods in the Magic of Big Data and AI', *Communication Monographs* 85, no. 1 (2 January 2018): 57–80, https://doi.org/10.1080/03637751.2017.1375130.
[119] boyd and Crawford, 'Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon'.

findings, despite scientists' conviction that their work engages exclusively with hard facts.[120] Accordingly, the development of AI and its corollary technologies is mediated by wavering *'temporalities, spatialities and materialities'*.[121] Individuals and institutions involved in AI development or usage inadvertently infuse AI systems and their outputs with subjective evaluations, such as cultural norms, strategic agendas, ideologies, career determinants, exigencies of professional practice, group dynamics, credibility and urgency judgements.[122]

At each step, AI developers undergo negotiations and power struggles to choose between options that reflect competing values and interests in order to shape the final product.[123] Upon selecting these options, their corresponding values and interests are embedded in AI. Under Friedman and Nissenbaum's taxonomy, social institutions, practices, and attitudes factored into the programming process become vehicles of 'preexisting bias'.[124] Beyond that, the AI system is shaped by programming tools, platforms, devices, and data ontologies alongside other technical considerations, which, by the same taxonomy, cause 'technical bias'.[125] Afterwards, the interaction between human operators and the AI system informs the formers' interpretation of the outputs produced by the latter and, in the other direction, opens up new avenues for 'emergent bias' to percolate into the ML model.[126]

Therefore, any assumption of Big Data as literal data (Latin, *datum*: given) and of AI developers as revealing hitherto concealed facts and truths is ill-founded. Rather than being a 'blank slate' mirroring back to society what is true, accurate, and value-neutral, AI systems are theory-laden and value-laden artefacts. As insinuated in Dwork and Mulligan's claim that *'the reality is a far messier mix of technical and human curating',* human curation and interpretation holds a strongly influential role throughout AI development.[127] Even further, as noticeable in the introductory real-life incidents, it imbues AI with biases on the following levels: biased selection of objectives; biased training data; proxy discrimination; threats to fairness and equality; inscrutable processing model; and intentional discrimination.

---

[120] Bruno Latour, Steve Woolgar, and Jonas Salk, *Laboratory Life: The Construction of Scientific Facts* (Princeton, United States: Princeton University Press, 1986), http://ebookcentral.proquest.com/lib/kcl/detail.action?docID=1144731.
[121] boyd and Crawford, 'Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon'.
[122] Elish and boyd, 'Situating Methods in the Magic of Big Data and AI'; Latour, Woolgar, and Salk, *Laboratory Life*, 135.
[123] Forsythe, 'Engineering Knowledge'.
[124] Batya Friedman and Helen Nissenbaum, 'Bias in Computer Systems', *ACM Transactions on Information Systems* 14, no. 3 (1 July 1996): 330–47, https://doi.org/10.1145/230538.230561.
[125] Friedman and Nissenbaum.
[126] Friedman and Nissenbaum.
[127] Martin, 'Ethical Implications and Accountability of Algorithms'.

2.3 Biased selection of objectives

AI is often utilised to determine, for instance, one's employability, creditworthiness, or likelihood of recidivism. As these outcomes reflect what human operators aim at finding, they are called target variables.[128] The possible values of each target variable, e.g. high, sufficient, low, or poor in the case of creditworthiness, constitute mutually exclusive categories called class labels.[129] The target variable is not always specified to begin with, so its definition falls on AI developers. This means that they are tasked, firstly, with understanding the objectives and requirements of the product, as this is conceived by the organisation for which it is developed.[130] Secondly, they have to translate this real-life problem into a question composed of a target variable and the class labels it can undertake.[131] In this selection process, developers are guided by their own perceptions of the problem at stake and its possible solutions. Willingly or not, they may identify a target variable which is systematically correlated with specific social groups and will, thus, distinctively affect them. For example, when building a recruitment AI to assess the best possible hires, if the developer selects as a target variable for the AI to identify candidates with the most accumulated, uninterrupted years of full-time work experience, it is highly probable that the AI system will negatively affect young people, whose mere quantity of work experience cannot exceed that of middle-aged candidates, and women, who represent the largest portion of part-time employees or take maternity leaves from work.[132]

Additionally, many target variables are not expressed in binary class labels, meaning that their outputs cannot be divided into 'yes' or 'no', but in scales or degrees, as in the case of creditworthiness. In this scenario, developers often create from scratch their own class labels.[133] There is no objective way, though, to determine, for instance, how many loan repayments one has to miss in order to be classified in the low class of creditworthiness or the poor one. Seemingly neutral categories have sociopolitical underpinnings, whereas others turn up to be overly fluid.[134] Even the mere practice of classifying individuals reduces them into oversimplified, quantified categories and groups. Such categorisations of individuals create an

[128] Barocas and Selbst, 'Big Data's Disparate Impact'.
[129] Barocas and Selbst.
[130] Barocas and Selbst.
[131] Barocas and Selbst.
[132] 'Part-Time Employment Rate' (OECD), accessed 16 November 2018, https://doi.org/10.1787/f2ad596c-en; 'Temporary Employment' (OECD), accessed 16 November 2018, https://doi.org/10.1787/75589b8a-en.
[133] Barocas and Selbst, 'Big Data's Disparate Impact'.
[134] Elish and boyd, 'Situating Methods in the Magic of Big Data and AI'.

illusion of uniformity among them, overlooking the nuanced, diverse identities of each person and deterring the exploration of alternatives.[135] As another form of power, algorithms set narrow boundaries for how people should behave and which are the possible courses of action, whilst anything that does not properly fit into their defined categories is discarded to a 'residual' one.[136] Although AI systems fare better when analysing consistent, predictable categories of individuals, human beings do not–and should not be expected to–conform to such disciplined ideals.[137] It is automation that should accommodate humans, not the other way round.

Indicatively, iBorderCtrl succumbs to this error by classifying facial expressions into lying or not lying, as if human reactions were so clearly defined or anticipated and in a way evoking pseudoscientific physiognomic endeavours of the past. Therefore, even from the commencement of its development, selecting the goals of AI reflects subjective, to a certain extent arbitrary, judgements.

2.4 Biased training data

The main pathway through which bias is introduced into AI are training data, viz. data which serve as learning examples for the ML model. The problem of biased training data is divided into two sub-cases: incorrect handling and historical data.

2.4.1 Incorrect handling

In the first sub-case, bias occurs through the handling of training data. As seen in Chapter 1, in supervised ML, a human agent assigns labels to the input data and the correct or not outputs of the model. Except for cases wherein human agents deliberately label data in overtly discriminatory ways, this may happen unwittingly because of implicit biases.

Indicatively, the prevailing computing culture, which rewards aggressive competition and detachment from computational objects, has failed women and thereby rendered the field overwhelmingly male-dominated.[138] Feminist scholars have long demonstrated that this gender

[135] Mike Ananny, 'Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness', *Science, Technology, & Human Values* 41, no. 1 (January 2016): 93–117, https://doi.org/10.1177/0162243915606523.
[136] Ananny.
[137] Ananny.
[138] Sherry Turkle and Seymour Papert, 'Epistemological Pluralism: Styles and Voices within the Computer Culture', *Signs* 16, no. 1 (1990): 128–57.

divide in the technological community profoundly influences the design and functions of technological products.[139] Which realities and problems are taken into account is eventually determined by the subject of deliberation, so in the case of technological artefacts it turns out that male experiences are placed at the forefront.[140] This implies that, for example, a developer might label the same job performance as worse for female than male applicants or that he might overlook including samples of female voices when training a speech recognition AI.[141]

In the same vein, less examined in the literature are AI biases caused by ableist perspectives. If developers of AI systems such as self-driving cars take their bodily constitution for granted and prioritise bodies similar to theirs, their algorithms will take decisions based on ableist assumptions and, for example, will not recognise people in wheelchairs in the streets.[142]

Creator's influence is similarly determined by disciplinary background. In general, social scientists gather and study data while being reflexive to the processes and environments in which they are generated. More so when social scientists are employed in decision-making positions, a context-sensitive viewpoint helps in discerning and articulating nuanced decisions or policies.[143] Contrarily, AI developers do not usually have domain expertise over the context from which their data are drawn.[144] Alternatively, they have no access to the original basis on which the training data were collected, as they often procure these from third parties, mainly data brokers, who are admittedly nonchalant about the contextualisation of data. As a result, it is possible to confuse Big Data as being 'whole data' and ignore that their collection is imperfect. This casts doubts on the legitimacy of developers as knowledge producers and, through their AI systems, decision-makers, and builds epistemic hierarchies around 'who can read the numbers'. Ultimately, disciplinary silos are erected between STEM (science, technology, engineering, mathematics) professionals and social scientists.[145]

---

[139] Judy Wajcman, 'Gender and Technology', in *International Encyclopedia of the Social & Behavioral Sciences*, ed. Neil J. Smelser and Paul B. Baltes, 1st ed, vol. 9 (Amsterdam; New York: Elsevier, 2001), 5976–5979.
[140] Wajcman.
[141] Speech recognition repeatedly performs better for men than women: Rachael Tatman, 'Google's Speech Recognition Has a Gender Bias', 12 July 2016, https://makingnoiseandhearingthings.com/2016/07/12/googles-speech-recognition-has-a-gender-bias/.
[142] Karen Hao, 'Can You Make an AI That Isn't Ableist?', MIT Technology Review, accessed 29 November 2018, https://www.technologyreview.com/s/612489/can-you-make-an-ai-that-isnt-ableist/.
[143] FET Advisory Group, 'The Need to Integrate the Social Sciences and Humanities with Science and Engineering in Horizon 2020 and Beyond' (European Commission, December 2016), https://ec.europa.eu/digital-single-market/en/news/report-need-integrate-social-sciences-and-humanities-science-and-engineering-horizon-2020.
[144] Citron, 'Technological Due Process'.
[145] boyd and Crawford, 'Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon'.

The chances of identifying such human positionality decrease when more than one agents are involved, each carrying their own biases over a protracted period, as is the case with AI development. What is certain, though, is that once human biases have tainted the labelling phase, they will apply to all future cases in which the same labels will be used by the ML algorithm. Greater diversity in AI development teams would be beneficial to offsetting this positionality, as it would entail the inclusion of a broader array of experiences, critical perspectives, and backgrounds into the process. Indicatively, teams consisting of Big Data practitioners and ethnographers, given that both fields share immersion in data as their methodology, could capitalise on the complementarity of their skills to engage in collaborative discoveries of data instead of compartmentalised ones.[146] In this way, ethnographers could witness the interaction of research subjects with complex media platforms, whereas Big Data practitioners would perceive the qualitative dimensions of data, especially of missing or incomplete ones, and their social context.[147]

Even if individual data are sufficiently precise and untarnished by developers' biases, it must be examined whether different social groups are adequately represented in the sample. Sample selection bias is possible when a part of the population is misrepresented, in which case data for this population group are invalid.[148] Being part of a population which is underrepresented or overrepresented in the sample data-set leads to systematically differential treatment. This became evident in Google's Photos application, whose face recognition AI had probably trained on data-sets including accurate pictures of mainly white people. Generally, face recognition AI systems are among the ones mostly infected by such biases on the grounds of race and gender.[149] This is troubling for the forthcoming launch of iBorderCtrl, which also uses AI for face recognition. Being denied entrance to a country just because one's face is of different skin colour or shape than the faces at which AI was trained is a scenario not to be taken lightly.

In cases of misrepresentation, attention should be paid to which groups are included but also to which are excluded. Not all lives are evenly 'datafied', as there are people who, due to low digital literacy or financial and geographical reasons, do not leave enough digital footprints

[146] Heather Ford, 'Big Data and Small: Collaborations between Ethnographers and Data Scientists', *Big Data & Society* 1, no. 2 (10 July 2014): 1–3, https://doi.org/10.1177/2053951714544337.
[147] Ford.
[148] Bianca Zadrozny, 'Learning and Evaluating Classifiers under Sample Selection Bias', in *Twenty-First International Conference on Machine Learning* (ICML '04, Alberta, Canada: ACM Press, 2004), 114, https://doi.org/10.1145/1015330.1015425.
[149] Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification', in *Proceedings of Machine Learning Research*, vol. 81, 2018, 77–91.

to be easily identifiable by Big Data collections and ML algorithms.[150] Despite this, businesses engaged in data mining and analysis do not demonstrate the same rigour as social scientists collecting data for academic research.[151] Because of their primarily commercial objectives, data about individuals and groups who do not have purchasing habits traceable by Big Data and ML are not of priority to businesses, so they remain on the margins of these tools. As they will not be offered products, promotions, and facilities suitable to their characteristics, they will be economically disadvantaged in a manner similar to discriminatory redlining practices.[152] As more governments employ Big Data and ML in their functions, information about such individuals is not captured in the public sector either, which could hinder their full participation to civil and political life.[153] All in all, both at governmental and industry level, the preferences, habits, and needs of these individuals are disregarded. With a growing use of AI in decision procedures, these individuals are plagued with a new 'voicelessness' when it comes to decisions about the distribution of services and goods or public policy reforms.[154] Therefore, it is imperative for those left out of data-driven technologies not to receive differential treatment because of this very exclusion. As Lerman calls it, they deserve protection under a 'data antisubordination principle' or 'a right not to be forgotten'.[155]

2.4.2 Historical data

In the second sub-case of biased training data, the output is affected by historical bias, as happened with Amazon's recruitment AI. When for historical reasons previously successful candidates of an application process belong to a specific group, training an ML algorithm on these candidates' data reinforces the selection of the same group, even if these historical reasons no longer exist. Conventions and norms in our society change and so do the decisions considered acceptable at each period. Conversely, when the training data-set includes cases which have been historically influenced by biases, the ML algorithm, which learns from examples, cannot distinguish that these cases are undesired as opposing current norms. So, it

[150] Barocas and Selbst, 'Big Data's Disparate Impact'.
[151] Barocas and Selbst.
[152] Jonas Lerman, 'Big Data and Its Exclusions', *Stanford Law Review Online* 66 (3 September 2013): 55–63, https://doi.org/10.2139/ssrn.2293765. See the example of Amazon Prime: David Talbot, 'Amazon's Same-Day Delivery Service Reinforces Inequality', MIT Technology Review, accessed 11 November 2018, https://www.technologyreview.com/s/601328/amazon-prime-or-amazon-redline/.
[153] Lerman, 'Big Data and Its Exclusions'. See the example of Street Bump: Kate Crawford, 'The Hidden Biases in Big Data', *Harvard Business Review*, 1 April 2013, https://hbr.org/2013/04/the-hidden-biases-in-big-data.
[154] Lerman, 'Big Data and Its Exclusions'.
[155] Lerman.

treats them as valid examples and automatically reproduces them and their attendant biases.[156] With each decision taken in this way, present or past biases of our society are amplified and maintained in the future. This is an almost unavoidable error in the case of language, whose discriminatory semantics are so deeply ingrained that automated translation machines cannot distinguish and avoid them, thus perpetuate gendered, sexist phraseology.[157] Hence, data traced back in time are not necessarily less biased.

This means that, firstly, AI developers should be better informed on cases of historical bias and discrimination so as to be in a position to recognise such tendencies in their algorithms. As an exemplary step towards this direction, Leavy leverages perspectives drawn from gender theory to pinpoint instances of gender bias in texts so that AI developers identify and avoid them in ML training.[158] Secondly, human agents should be cautious in taking conclusions reached by ML at face value. Since such data-driven algorithms conclude what should happen in a present case based on their Big Data analysis of what has so far been happening, they are susceptible to the naturalistic fallacy, which instructs that deriving normative prescriptions (what 'ought to be') from descriptive premises (what 'is') is invalid.[159]

The provisional exposition highlights that AI decisions are as good as the data upon which AI was trained. Increases in the volume of data are actually more likely to exaggerate any incorporated bias than annihilate it. As a widely used in the technology industry adage states 'Garbage In, Garbage Out'. Relatedly, Ioannidis demonstrated that a sizeable part of research claims preserves dominant biases, resulting to the falsity of most published research, and presented with Chavalarias a classification of 235 biases across biomedical research.[160] On account of that, doubting the quality of data in research as a whole would not be far-fetched.

---

[156] Barocas and Selbst, 'Big Data's Disparate Impact'.
[157] Muneera Bano, 'Artificial Intelligence Is Demonstrating Gender Bias – and It's Our Fault', 25 July 2018, https://www.kcl.ac.uk/news/news-article.aspx?id=c97f7c12-ae02-4394-8f84-31ba4d56ddf7; Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan, 'Semantics Derived Automatically from Language Corpora Contain Human-like Biases', *Science* 356, no. 6334 (14 April 2017): 183–86, https://doi.org/10.1126/science.aal4230.
[158] Susan Leavy, 'Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning', in *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering*, GE '18 (New York, USA: ACM, 2018), 14–16, https://doi.org/10.1145/3195570.3195580.
[159] Tae Wan Kim, Thomas Donaldson, and John Hooker, 'Mimetic vs Anchored Value Alignment in Artificial Intelligence', *ArXiv:1810.11116 [Cs]*, 25 October 2018, http://arxiv.org/abs/1810.11116.
[160] John P. A. Ioannidis, 'Why Most Published Research Findings Are False', *PLOS Medicine* 2, no. 8 (30 August 2005): 696–701, https://doi.org/10.1371/journal.pmed.0020124; David Chavalarias and John P.A. Ioannidis, 'Science Mapping Analysis Characterizes 235 Biases in Biomedical Research', *Journal of Clinical Epidemiology* 63, no. 11 (November 2010): 1205–15, https://doi.org/10.1016/j.jclinepi.2009.12.011.

2.5 Proxy discrimination

The second level at which bias crops up is proxy discrimination. An ML model might be fed with detailed data inputs which are genuinely related to the output, without explicitly referring to any protected or sensitive attributes, such as race or gender. However, it might be revealed that the output, e.g. job performance or risk score, closely correlates with membership in a protected group. In this case, and in an effort to reach maximum predictive accuracy, the inputs are indirectly sorted according to attributes which should not be taken into consideration. This especially happens because of 'redundant encoding', when the algorithm is programmed to receive as input more than the strictly necessary features of data, and thereby membership in a protected social group ends up being encoded in them.[161] Because of this extended scope of data collection, information about one's group membership can be used as an indirect proxy for the output. For instance, if an insurance pricing algorithm finds a pattern associating drivers of sports cars with higher risks of accidents, its output will suggest a higher premium. If sports cars are also mostly owned by male drivers, the fact that someone is male is used as an indirect proxy, which boils down to them paying higher premiums than comparable female drivers.[162] Similarly, in the case of COMPAS, even if defendants' race is not among the data directly used as input, they are asked whether one of their parents was ever sentenced in jail or prison.[163] In the US context, this correlates with race, given the discriminatory drug laws and prosecutions of the 1980s.[164]

Relatedly, it is possible for ML algorithms to disclose novel, unforeseen connections among data belonging to the same data-set, upon repeated uses of the latter. This practice of trawling through data to extract every possible association is disparagingly referred to as 'data dredging'.[165] One of its distinct features is the paradigm shift from causation to correlation. Although it is a scientific staple that 'correlation does not imply causation' and cannot be informative regarding the underlying causes of phenomena, for proponents of Big Data analysis through ML, mere predicting based on massive amounts of data and their multifarious relations is satisfactorily reliable. What is more, this data-driven analysis can be used in almost

[161] Barocas and Selbst, 'Big Data's Disparate Impact'.
[162] Philipp Hacker, 'Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU Law', *Common Market Law Review* 55, no. 4 (1 August 2018): 1143–85.
[163] Angwin et al., 'Machine Bias'.
[164] Kenneth Nunn, 'Race, Crime and the Pool of Surplus Criminality: Or Why the "War on Drugs" Was a "War on Blacks"', *UF Law Faculty Publications*, 1 January 2002, https://scholarship.law.ufl.edu/facultypub/107.
[165] David Jensen, 'Data Snooping, Dredging and Fishing: The Dark Side of Data Mining A SIGKDD99 Panel Report', *SIGKDD Explorations* 1, no. 2 (January 2000): 52–54.

any context, whereas alternatives might be costly, completely unavailable, and equally fraught with errors. Therefore, there seems to be no need to theorise, unpack, or challenge the identified correlations. In our Petabyte Age, semantic and causal analyses have turned obsolete.[166]

Quoting Calude and Longo:

> *'[…] ''co-relation'' denotes phenomena that relate covariantly, that is, they vary while preserving proximity of values according to a pre-given measure. ''Co-relation'' is essentially ''co-incidence'', that is, things that occur together. Correlations can be useful because of their potential predictive power: use or act on the value of one variable to predict or modify the value of the other.'*[167]

Notwithstanding their predictive abilities, correlations should not be considered sufficiently credible evidence to justify decisions and actions. Compared to causal connections, correlations are inconclusive and irreproducible, which diminishes their validity. The research of Caruana et al., who applied ML to predict which pneumonia patients were most at risk, exemplifies such precarious correlations. Surprisingly, one of the conclusions generated by ML was that asthmatic patients were less at risk of dying of pneumonia.[168] The ML model had rightly identified that as a result these patients had lower death rates, but in fact they were of high, not low, risk.[169] What the model did not capture and accordingly did not reflect was that this happened because asthmatic patients received earlier and more intensive healthcare due to their previous medical history.[170] Mindlessly relying on correlations like this to prioritise hospitalisation or funnel medical benefits could be life-threatening for patients.[171] Often enough, these correlations reveal patterns which do not even exist, comparably to what the German psychiatrist Conrad coined as apophenia, namely the propensity to perceive illusory connections and meaningfulness among random phenomena.[172] Therefore, although Big Data and ML are sufficient to extrapolate patterns, these are exceedingly infused with spurious, randomly generated correlations; as such, they are unreliable sources of information.[173]

---

[166] Anderson, 'The End of Theory'.
[167] Cristian S. Calude and Giuseppe Longo, 'The Deluge of Spurious Correlations in Big Data', *Foundations of Science* 22, no. 3 (September 2017): 595–612, https://doi.org/10.1007/s10699-016-9489-4.
[168] Rich Caruana et al., 'Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission', in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '15, Sydney, Australia: ACM Press, 2015), 1721–30, https://doi.org/10.1145/2783258.2788613.
[169] Caruana et al.
[170] Caruana et al.
[171] See the examples of healthcare in Arkansas, US: Colin Lecher, 'A Healthcare Algorithm Started Cutting Care, and No One Knew Why', The Verge, 21 March 2018, https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy.
[172] boyd and Crawford, 'Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon'.
[173] Calude and Longo, 'The Deluge of Spurious Correlations in Big Data'.

2.6 Threats to fairness and equality

As a next step, AI systems are evaluated based on their morally significant effects. For Binns, the opposition of AI to fairness lies on egalitarian yardsticks.[174] The fundamental moral principle of egalitarianism is equal treatment of all people and, according to some of its strands, equal distribution of valuable resources.[175] AI decision-making allocates individuals to groups, classes, or constructed profiles. Despite their real-life behaviour, the similarity of their data with those of other group members is deemed sufficient to have information about their allocated group, class, or profile unvaryingly applied to them and linked with positive or negative effects related to the distribution of finite social goods, such as education, medical care, housing, or employment. Thus, AI systems act as gatekeepers, granting or denying valuable opportunities and access to welfare.

In particular, luck egalitarianism posits that inequalities resulting from luck are not morally permissible as opposed to those resulting from individuals' free and informed choices.[176] This implies that whenever AI systems take into account data on an individual which are attributed to luck and not choice, their outputs are morally wrong. Following the doctrine of luck egalitarianism, the decisions reached by COMPAS were morally impermissible, as it considered variables beyond the defendants' control, such as being born into a neighbourhood with higher crime rates or by parents with a criminal record, and accordingly caused the differential treatment of a protected social group.

Deontic egalitarianism explains why the emergence of historical bias, touched upon in part 2.4.2, is ethically problematic. Deontic egalitarians evaluate not only an unequal state of affairs as it is, but the course of its establishment from an interwoven historical, economic, and sociological perspective.[177] Relatedly, they are concerned with the attribution of responsibility for the emergence and elimination of inequalities. When COMPAS predicts that African Americans are more likely to commit crimes because its historical data show that this used to be the case, it disregards the racial profiling, negative stereotypes, and historical conditions behind such statistics. When Amazon's recruitment AI predicts that women are not likely to be successful employees because men have so far been its most successful ones, it disregards

[174] Reuben Binns, 'Fairness in Machine Learning: Lessons from Political Philosophy', in *Proceedings of Machine Learning Research*, vol. 81, 2017, 1–11, http://arxiv.org/abs/1712.03586.
[175] Binns.
[176] Binns.
[177] Binns.

the socioeconomic factors which delayed women's entry to the workforce and have kept them under glass ceilings. When Google Photos labels images of African Americans as gorillas, it disregards the cultural weight of such characterisations and their historical usage to dehumanise black people. As long as ML algorithms do not include such contextual considerations in their model, they admit to ethical shortcomings.

Within the framework of egalitarian fairness, equal distribution is not necessarily conceived as direct allocation of benefits and harms to individuals. It is likewise understood as equal representation of different social groups.[178] As previously noted, biased training data lead to inaccurate representations of society. An illustrative example is found in the research of Bolukbasi et al., who demonstrated how search engine algorithms group words in gender-biased categories. Job titles such as boss or financier are allocated in the category of male job positions, whereas receptionist or housekeeper are allocated in female job positions.[179] If an AI is trained on texts containing stereotypical language, it will inevitably reproduce this biased representation in its outputs. Such representational harms are interdependent with unfair outcomes and, especially when decision-makers act upon them, cause allocative harms.[180] This is easily understood if we think of resources such as job vacancies, which are allocated partly based on people's perceived representation e.g. in online search results. In the long run, discriminatory representations turn to self-fulfilling prophecies, not merely indicating which individuals are 'wheat' and which 'chaff' but actively contributing to their being and self-identifying as such.[181] Representing certain individuals or groups as unworthy of social goods stigmatises them and reinforces their subordination, thus perpetuating oppressive cycles. In sum, equal distribution and representation across social groups comprises the egalitarian account of fairness.

All these considered, the Rawlsian demand for any socioeconomic discrepancies to be in favour of the underprivileged members of polity ('the difference principle') is not fulfilled,

[178] Binns.
[179] Tolga Bolukbasi et al., 'Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings', *ArXiv:1607.06520 [Cs, Stat]*, 21 July 2016, http://arxiv.org/abs/1607.06520. See also: Latanya Sweeney, 'Discrimination in Online Ad Delivery', *SSRN Electronic Journal*, 28 January 2013, https://doi.org/10.2139/ssrn.2208240.
[180] Roel Dobbe et al., 'A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics', in *2018 Workshop on Fairness, Accountability, and Transparency in Machine Learning* (ICML 2018, Stockholm, Sweden, 2018), http://arxiv.org/abs/1807.00553.
[181] Danielle Keats Citron and Frank Pasquale, 'The Scored Society: Due Process for Automated Predictions', *Washington Law Review*, University of Maryland Legal Studies Research Paper No. 2014-8, 89 (2014): 1–33.

thus fairness is not achieved.[182] The fact that such unfair practices seem to originate not from a human but an artificial entity does not make the experience of discrimination and its effects less real. On the contrary, AI engenders the threat of discriminating in unfamiliar, scalable ways, yet often with long-lasting egregious effects.

2.7 Inscrutable processing model

In traditional cases of human decision-making, it is expected that the connection between the conclusion and the data from which the conclusion was surmised is accessible to affected parties so that they can examine the justification of the decision and challenge it. Accordingly, remedying AI biases presupposes the ability to diagnose them.[183] Nonetheless, recognising subjective and arbitrary elements in an ML algorithm is so subtle and observer-dependent that victims of discrimination might not even know that they are being discriminated against.

On the one hand, individuals affected by data-driven decisions usually have limited knowledge of the full extent, provenance, and quality of their used data. Concurrently, AI programming languages and tools remain unfathomable to the general public, who thus cannot access, read, or assess the source code of AI-enabled decisions, and accessible only to a fraction of specialists.[184] Placing in juxtaposition the infrastructure, resources, and influence of eminent organisations or technological corporations (nicknamed as 'Big Tech') engaged in AI and the individuals whose data are harvested for AI development and whose life is affected by decision-making algorithms reveals a vast epistemic inequality and power imbalance. Therefore, lay people are impeded, at a first level, from comprehending how AI processes their data and, at a second level, from exerting oversight on whether the decisions reached are biased or not.

On the other hand, what is distinct about ML algorithms is that full comprehension of their inner workings is unattainable even for domain experts. Before the emergence of a problematic case and as long as an AI system performs well during its training, developers are unaware of latent biases. It is only when ML algorithms fail, that their existing flaws become overt. Even then, and despite their technical expertise, AI developers cannot understand the process followed by ML algorithms due to their self-learning abilities and the aggregation of

[182] Leif Wenar, 'John Rawls', in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta, Spring 2017 (Metaphysics Research Lab, Stanford University, 2017), https://plato.stanford.edu/archives/spr2017/entries/rawls/.
[183] Friedman and Nissenbaum, 'Bias in Computer Systems'.
[184] Burrell, 'How the Machine "Thinks"'.

high-dimensional data. AI systems are designed with effective prediction, not interpretability, as their rewarded function, which means that their efficiency comes at the cost of understanding.[185]

More precisely, ML algorithms, especially unsupervised ones, adapt their models as they receive inputs from their environment. This implies a dynamic modification of their inner structure according to the feedback they receive. Specifically, when deployed in contexts with human interaction, the ML model adjusts its behaviour according to the new data that users give as inputs. Any human biases included in these new data are learnt by the ML model, resulting in 'emergent bias'.[186] This is what happened with Microsoft's Tay, that learned racism and sexism by interacting with Twitter users. Because of the fluidity, complexity, and speed of this process, not only the final outputs of ML models are unpredictable by design but the ways in which they used the initial inputs are unintelligible by AI developers, let alone affected individuals. This inability to interpret AI does not demonstrate an inadequacy in terms of expertise or practical resources. It refers to the extent to which *'a human can articulate the trained model or rationale of a particular decision, for instance by explaining the influence of particular inputs or attributes'*.[187] Because of the lack of interpretability, it cannot be confirmed whether a questionable case is just a one-off problem or indicative of a structurally biased model.

The endemic opacity of AI thwarts efforts to inspect and monitor it, resulting to its description as a highly problematic but incontrovertible 'black box', which turns inputs to outputs without any visible clue of the intermediary steps. As early as 1979, Latour and Woolgar defined black-boxing in science as *'rendering items of knowledge distinct from the circumstances of their creation'*.[188] More recently and with the aim of illustrating the powerful implications of gaps in knowledge, Pasquale draws parallels between the opaque inner workings of algorithms and Adam Smith's 'invisible hand' of the market.[189] In his own words:

> *'The term "black box" is a useful metaphor for doing so, given its own dual meaning. It can refer to a recording device, like the data-monitoring systems in planes, trains, and cars. Or it can mean a system whose workings are mysterious; we can observe its inputs and outputs, but we cannot tell how one becomes the other. We face these two meanings daily: tracked ever more*

[185] Elish and boyd, 'Situating Methods in the Magic of Big Data and AI'.
[186] Dobbe et al., 'A Broader View on Bias in Automated Decision-Making'.
[187] Mittelstadt et al., 'The Ethics of Algorithms'.
[188] Latour, Woolgar, and Salk, *Laboratory Life*, 219.
[189] Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Cambridge, United States: Harvard University Press, 2015), 2, http://ebookcentral.proquest.com/lib/kcl/detail.action?docID=3301535.

> *closely by firms and government, we have no clear idea of just how far much of this information can travel, how it is used, or its consequences.'*[190]

Eventually, there is an information asymmetry wider than the one between lay individuals and AI developers: that of human agents and machines in general. This asymmetry hinders the effective exercise of human agency, as without intelligible interpretations individuals cannot properly assess the risks of data processing and guide their actions.[191] This weakening of human agency expands to the public sphere, wherein human participation is constrained due to a growing reliance on purportedly comprehensive and impartial algorithms, creating what Danaher dubs a system of 'algocracy'.[192]

2.8 Intentional discrimination

All the preceding sorts of biased and discriminatory decisions may occur unwillingly, but equally enable developers to disguise their deliberateness behind allegedly neutral models. In pursuit of ways to avoid compliance with law or conceal violated regulations, discriminatory patterns, and other illegitimate activities, corporations conceivably fabricate and invoke the authoritative appeal as well as the opacity of critical algorithms.[193]

In particular, developers of AI systems could intentionally shape the sample data-set in a biased way so that its outputs discriminate in favour or against a social group, or they could insist on the validity of historical data albeit cognisant of their partiality. Exploiting the possibilities of proxy discrimination, operators of AI systems can distinguish which individuals belong to specific social groups and alternate their treatment, even if the individuals themselves have not explicitly given such data as input. In practice, by leveraging ML and Big Data, self-serving organisations can bypass long-standing legislation and codes of conduct prohibiting discrimination of individuals based on their membership in social groups. On the pretext that this is an impartial, yet inscrutable even for the developers themselves technological process, natural and legal persons are able to mask entrenched or recent ways of discrimination and protect themselves against public backlash.

---

[190] Pasquale, 3.
[191] Mittelstadt et al., 'The Ethics of Algorithms'.
[192] John Danaher, 'The Threat of Algocracy: Reality, Resistance and Accommodation', *Philosophy & Technology* 29, no. 3 (September 2016): 245–68, https://doi.org/10.1007/s13347-015-0211-1.
[193] Pasquale, *The Black Box Society*.

2.9 Conclusion

The usefulness of AI is not denied; neither should we forget that human decision-making is rife with biases. Yet, the gap between the development of such technologies and our ability to apprehend and control their morally significant aspects needs to be bridged, if laws are about to regulate them and society to heavily rely on them. Specifically, demands for epistemically sound and fair decisions should be stricter, proportional to the role of AI in decision-making and the societal significance of the decisions at stake.

In this direction, the design of AI needs to be revised and become value-sensitive, lifting the lid on values embodied or overlooked during its development and facilitating the incorporation of ethically desirable ones such as fairness and equality.[194] Shortly, AI should no longer be optimised solely for cost saving and efficiency but for human wellbeing, too. On their part, human agents involved in the process should engage in critical self-reflection and interdisciplinary coalitions in order to better identify their own biases and address them with a cross-fertilisation of multiple epistemic viewpoints.

As long as these are not achieved, the shift to a data fundamentalism, which regards the analysis of Big Data through AI as unequivocally legitimate means of decision-making and action-guiding, negatively affects humans at an individual and group level. Hence, it appears safer to keep using them and their correlative powers but feed their insights as inputs into a causal model for understanding.[195] In this way, AI will hold an advisory, less definite role, especially in contexts vital to human rights, allowing human agents to engage in vigilant knowledge production and decision-making. Eventually, the epistemic and normative limitations of AI should be acknowledged, leading to its use for facilitating instead of replacing human decision-making.

In the next Chapter, these ethical deficiencies of AI systems help us understand why their governance under data protection laws is a daunting exercise.

[194] Dobbe et al., 'A Broader View on Bias in Automated Decision-Making'.
[195] Dobbe et al.

Chapter 3: The legal challenges of AI

3.1 Introduction

The discriminatory biases of Chapter 2 are of ethical but also legal interest, as AI is increasingly used to take or facilitate decisions with wide-sweeping legal effects. It is already ethically challenging when AI produces biased outcomes, for instance in recognising faces in photos, but such biases preponderate if they creep into self-driving vehicles, lethal autonomous weapons, university admissions, social benefits, or even citizen scoring. Nonetheless, the AI market has insufficient incentives to self-regulate, as this would demand the allocation of more human and financial resources and would, thus, be unattractive from a cost-benefit perspective. As a result, notwithstanding the media furore sparked every time an AI system proves to be biased, these problems are persisting.

Whether EU policies address the foregoing ethical concerns is now ripe for exploration. In response to the turmoil of biased AI, the EU policy armamentarium comprises, on the one hand, data protection legislation enforceable in AI-powered data processing and, on the other hand, soft law targeted to AI technologies.

3.2 Current landscape

The European Commission regularly acknowledges the biases that abound in AI decision-making and calls for their mitigation.[196] In the wake of such challenges, both the Commission and the European Parliament have stressed the importance of ensuring compliance of the AI field with the GDPR. On its part, the Commission has called the attention of national data protection authorities and the European Data Protection Board to the matter.[197] Moreover, it promised to leverage the work of the European Consumer Consultative Group and the

[196] Secretariat-General, 'A Connected Digital Single Market for All', Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on the Mid-Term Review on the implementation of the Digital Single Market Strategy (Brussels: European Commission, 5 October 2017), https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1511969120337&uri=CELEX:52017DC0228&print=true; Directorate-General for Communications Networks, Content and Technology, 'Artificial Intelligence for Europe'; Directorate-General for Communications Networks, Content and Technology, 'Coordinated Plan on Artificial Intelligence', Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions (Brussels: European Commission, 7 December 2018), https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1544192414694&uri=COM:2018:795:FIN.

[197] Directorate-General for Communications Networks, Content and Technology, 'Artificial Intelligence for Europe'.

European Data Protection Board to equip consumer organisations and data protection authorities with knowledge on AI and its applications.[198]

The WP29 also highlighted the applicability of GDPR to AI, ML, and Big Data. Advancements in these technologies ease the creation of profiles, which are subsequently used for automated decision-making.[199] The effects on individuals' rights and freedoms are acute, for instance, when they are denied access to employment, credit, and insurance or they become targets of deceptive advertising.[200]

Despite the value of GDPR–as well as its counterpart Police Directive for crime-related data–to AI, the European Parliament illustrated in its Resolution that the interconnectedness of emerging technologies and their ability to function with varying degrees of autonomy necessitates the regulation of new data protection issues.[201] Although ML has catalysed data analysis, it severely affects final decisions of consumer, business, or authoritative nature and challenges non-discrimination, due process, transparency, and explainability demands.[202] Similarly, the Commission stated that it would consider legislative adjustments based on new developments in AI.[203] Speaking more specifically in its Communication *On the Road to Automated Mobility: An EU Strategy for Mobility of the Future*, it claimed that, although the existing data protection legislation is internationally acclaimed for its high standards and facilitates technological progress without compromising EU values, it is imperative to introduce updated regulations due to advancements in automation.[204] Of the same attitude is the EESC, which suggested the consolidation of a clear, harmonised, and mandatory legal framework to address the risks of data-driven AI, including discrimination.[205]

A legislation kept abreast with AI development would inspire individuals and businesses to place confidence in the technology they use within a legal environment that

[198] Directorate-General for Communications Networks, Content and Technology.
[199] Article 29 Data Protection Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679', 3 October 2017, http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053.
[200] Article 29 Data Protection Working Party.
[201] European Parliament et al., 'Civil Law Rules on Robotics'.
[202] European Parliament et al.
[203] Secretariat-General, 'A Connected Digital Single Market for All'.
[204] Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs, 'On the Road to Automated Mobility: An EU Strategy for Mobility of the Future', Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee, the Committee of the Regions (Brussels: European Commission, 17 May 2018), https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1541878088823&uri=CELEX:52018DC0283.
[205] Giuseppe Guerini, 'Artificial Intelligence for Europe (Communication)', Opinion (Brussels: European Economic and Social Committee, 19 September 2018), https://www.eesc.europa.eu/en/our-work/opinions-information-reports/opinions/artificial-intelligence-europe-communication.

foreseeably and effectively protects EU values, rights, and freedoms.[206] Regardless of new technologies emerging, the deep-seated principles of necessity and proportionality along with the EU's commitment to justice, equality, dignity, and non-discrimination, as indicatively enshrined in Article 2 of the Treaty on European Union (TEU), need to be defended.[207] Human dignity and autonomy have also been the EGE's cynosure: the former restricts automated classifications of individuals as opposed to self-determination; the latter points out humans' ability to select if and under what circumstances decision-making will be delegated to AI.[208] Autonomy is further linked to transparency and predictability, as their absence deprives individuals of the ability to effectively interrupt or terminate an AI system.[209]

To that end, under its Digital Single Market Strategy, the Commission has inaugurated the Algorithmic Awareness Building project to collect evidence and assist policy-making on the challenges of automated decisions, including biases and discrimination.[210] In parallel, it has appointed a High-Level Expert Group on Artificial Intelligence (AI HLEG), which holds an advisory role in the Commission's AI strategy and is tasked with proposing AI ethics guidelines and general policies.[211] By the same token, it introduced the European AI Alliance, a multi-stakeholder initiative to engage with and offer insights to the AI HLEG.[212]

Moreover, the Commission pledged to mediate intra-EU conversations on AI.[213] In 2018, Member States signed a *Declaration of cooperation on Artificial Intelligence (AI)* and agreed on formulating a legal and ethical framework that will respect the EU's fundamental rights and values.[214] Within the Digitising European Industry framework, the Commission set in motion the collaboration among signatory states in the form of a High-Level Forum of

[206] Directorate-General for Communications Networks, Content and Technology, 'Artificial Intelligence for Europe'.
[207] European Parliament et al., 'Civil Law Rules on Robotics'.
[208] European Group on Ethics in Science and New Technologies, 'Artificial Intelligence, Robotics and "Autonomous" Systems'.
[209] European Group on Ethics in Science and New Technologies.
[210] Directorate-General for Communications Networks, Content and Technology, 'Artificial Intelligence for Europe'.
[211] European Commission, 'Artificial Intelligence: Commission Kicks off Work on Marrying Cutting-Edge Technology and Ethical Standards', 3 September 2018, http://europa.eu/rapid/press-release_IP-18-1381_en.htm.
[212] Directorate-General for Communications Networks, Content and Technology, 'Artificial Intelligence for Europe'.
[213] Directorate-General for Communications Networks, Content and Technology.
[214] Séverine Waterbley et al., 'Cooperation on Artificial Intelligence', Declaration, 4 October 2018, https://ec.europa.eu/digital-single-market/en/news/eu-member-states-sign-cooperate-artificial-intelligence.

Member States and an AI Forum in Finland.[215] Eventually, this collaboration between Member States and the Commission transformed into a *Coordinated Plan on Artificial Intelligence*.[216]

Having mapped the current landscape, the remaining parts examine legal provisions which apply before, during, and after the AI-enabled biased processing of data. By assessing the suitability of the GDPR for biased AI, this Chapter unpacks the tensions created and vindicates the chorus of EU voices considering the current legal regime insufficient.

3.3 Definitions under Article 4 GDPR

For the purposes of the Regulation, Article 4 (1) defines personal data as *'any information relating to an identified or identifiable natural person ('data subject')'*. Data such as name, identification number, location data, and online identifiers directly single out individuals. Indirect identification of persons occurs through information linked to their physical, physiological, genetic, mental, economic, cultural, or social identity, or combinations of these.

Any operation or set thereof performed on personal data is processing, according to Article 4 (2), including: *collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction*. Whichever natural or legal person, on its own or with others, determines the means and purposes of processing is a data controller (Article 4 (7)).

These definitions are necessary to affirm whether biased AI decisions fall under the scope of the GDPR, which covers personal data processing by automated means (Article 2 (1)). The first condition in the definition of AI, seen in Chapter 1, is its composition of algorithms which analyse data. As such, AI systems are automated means, in contrast to manual. Personal data processed by AI are directly provided by data subjects (e.g. through questionnaires), produced through observation of data subjects (e.g. location data through the use of applications), derived or inferred, that is, produced by previous profiling (e.g. one's credit score).[217] The UK's Information Commissioner's Office (ICO) distinguishes between derived

[215] 'Commissioner Gabriel Hosted First High-Level Forum of Member States on Digitalisation of Industry and Artificial Intelligence', Digital Single Market, accessed 24 November 2018, https://ec.europa.eu/digital-single-market/en/blogposts/commissioner-gabriel-hosted-first-high-level-forum-member-states-digitalisation-industry; 'AI Forum 2018 in Finland', Digital Single Market, accessed 24 November 2018, https://ec.europa.eu/digital-single-market/en/news/ai-forum-2018-finland.
[216] Directorate-General for Communications Networks, Content and Technology, 'Coordinated Plan on Artificial Intelligence', 7 December 2018.
[217] Article 29 Data Protection Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679'.

data as produced in relatively simple ways and inferred ones as produced by more complex, precarious analyses based on correlation and classification.[218]

Data processing through AI occurs mainly in two stages: during the training of AI, wherein training data may include personal data, and upon its deployment in decision-making, wherein personal data are used as input but may also be derived from the algorithm as output. In identifying the data controller of an AI system, the lines are often blurry, given, firstly, the large number of individuals and companies involved in its development and usage and, secondly, that in most cases no human agent is fully aware of its internal processes. However, at least to the extent that they determine the objectives, basic functions, and types of data processed by the AI system, the role of data controllers is assigned to AI developers during its training and to its operators/users during its deployment.

3.4 Data processing principles under Article 5 GDPR

Article 5 lays down the principles which shall apply to data processing. The principle of purpose limitation in Article 5 (1) b) justifies the collection of personal data for specified, explicit, and legitimate purposes and prohibits further processing for purposes incompatible with these. The principle of data minimisation in Article 5 (1) c) postulates the use of personal data to the extent that they are relevant and necessary to the purposes of processing. On the basis of the principle of accuracy in Article 5 (1) d), personal data should be accurate and, if necessary, updated, whereas inaccurate ones should be promptly erased or rectified. Finally, the principle of storage limitation in Article 5 (1) e) requires keeping personal data only for the period necessary for processing. Article 5 (1) a) contains the principle of lawful, fair, and transparent processing, which merits separate examination in part 3.7.[219] In a nutshell, these principles intend to ensure that personal information is processed without compromising individuals' protection. The Police Directive adopts the said principles of purpose limitation, minimisation, accuracy, and storage limitation regarding crime-related data in Article 4 (1) b), c), d), and e) respectively.

In paragraph 2 of Article 5, the GDPR introduces an explicit principle of accountability, according to which data controllers are bound to be responsible for and demonstrate compliance with the abovementioned principles. Especially in the context of discriminatory,

[218] Information Commissioner's Office, 'Big Data, Artificial Intelligence, Machine Learning and Data Protection', 1 March 2017, https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf.
[219] The principle of integrity and confidentiality in Article 5 (1) f) is not of relevance here.

erroneous, or unjustified ML algorithms, this provision could be interpreted as substantiating the enduring demand for 'algorithmic accountability'.[220] Yet, as we will see below, it is difficult for data controllers to reconcile AI with the principles of paragraph 1.

3.4.1 Principle of purpose limitation and AI

The principle of purpose limitation entails a two-fold requirement. Firstly, it demands from the outset a clear specification of the reasons and purposes of processing so that data subjects can grant their informed consent and control the usage of their data. Nonetheless, as seen earlier, ML algorithms process (Big) Data without necessarily forming in advance hypotheses. They serendipitously divulge patterns and correlations among data points, even regarding questions unimagined by controllers and data subjects. Delimiting the purpose of processing so broadly that it covers as many unexpected outcomes of processing as possible would oppose the requirement of a specific purpose. Thus, controllers cannot specify and explicitly state the purpose of processing in advance. Instead, only after such processing occurs will its purposes become epistemically accessible to them and consequently to data subjects. Hence, the inductive and dynamic character of AI contravenes this first requirement.

Secondly, the principle of purpose limitation mandates a compatibility test. Except for when data subjects' consent or national/EU law specifically allows it, each case of new processing is examined in terms of its compatibility with the purposes for which data were initially collected. Pursuant to Article 6 (4), this compatibility test relies on the relationship between the purposes of initial and intended data processing; the context of the initial data collection; the purposes of the intended further processing; the nature of data; the anticipated impact of the intended processing; and the safeguards applied by the controller. ML commonly reuses already gleaned and processed data-sets to reveal new correlations for purposes different than those of their original collection. On the basis thereof, if data controllers cannot verify compatibility between extant and intended processing, they should seek the data subjects' updated consent or alter the purpose of processing.

The WP29 demarcated two scenarios on weighing compatibility in relation to Big Data.[221] In the first case, wherein organisations use ML to descry tendencies in Big Data, the concept of functional separation comes into play, precluding the application of such data to

[220] Information Commissioner's Office, 'Big Data, Artificial Intelligence, Machine Learning and Data Protection'.
[221] Article 29 Data Protection Working Party, 'Opinion 03/2013 on Purpose Limitation', 2 April 2013, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf.

back measures or decisions affecting data subjects.[222] Compatibility will be confirmed only if such functional separation has been warranted. In the second case, wherein organisations parse Big Data to zero in on individuals' predilections and activities and accordingly take measures or decisions about them, compatibility will be reckoned only if *'free, specific, informed and unambiguous 'opt-in' consent'* has been conferred.[223] Yet, for AI systems used as decision aids, the prerequisite of specific and informed consent to affirm the compatibility of forthcoming processing does not square with their unpredictability by design and the inability to in advance specify the purposes of processing. Therefore, AI decision-making fails this second requirement of purpose limitation, too.

A possible diversion could occur through Article 5 (1) b) referring to Article 89 (1) on archiving in the public interest, scientific or historical research, and statistical analyses and establishing their a priori compatibility with the initial purposes of processing. In that sense, AI development could be perceived as scientific research, as according to the broad interpretation of Recital 159 this includes *'technological development and demonstration, fundamental research, applied research and privately funded research.'* Alternatively, including the deployment of AI in the realm of statistical analyses is rebutted by Recital 162 which forbids their use in decision-making. Given that Recitals are not legally binding and that the main Articles define neither scientific research nor statistical analysis, it remains to be seen whether AI will be subsumed under this provision.

3.4.2 Principle of data minimisation and AI

AI puts pressure on the principle of data minimisation, which obliges controllers to use only as many data as necessary and relevant to the purposes of processing. AI capitalises on large quantities of data, especially Big Data, in order to be effective. The goal is to glean as many data as are available and, if possible, all of them, as could be summarised by *'n=all'*, where 'n' refers to the sample size in statistics.[224] Beyond that, it is difficult to identify in advance the purposes of processing and, a fortiori, which data are strictly necessary and relevant to these purposes, as ML algorithms continuously adapt their models in interaction with their environment. Significantly, data aggregation beyond necessary limits is not uncommon, as 72% of businesses in the UK, France, and Germany admitted possessing data which, despite

[222] Article 29 Data Protection Working Party.
[223] Article 29 Data Protection Working Party.
[224] Viktor Mayer-Schönberger and Kenneth Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (Boston, MA: Houghton Mifflin Harcourt, 2013), 26.

their collection, ended up unused.[225] Apart from the amount of data, minimisation applies to their nature. Pseudonymisation and encryption are endorsed to make sure that a subject's identity is not disproportionately exposed; yet, such measures are not preferred as undermining the accuracy of AI outputs. Overall, the scale, variety, detail, and multiple sources of provenance of Big Data used by AI collide with data minimisation.

Although Microsoft's Tay was quickly terminated, if it had insisted on racist and sexist interactions with individuals, the principle of data minimisation would be a useful counterattack, since it was processing data fed by Twitter users, without distinguishing between those necessary or relevant and those not. Similarly, COMPAS is at odds with data minimisation. The fact that its ML algorithm receives as input a total of 137 variables, including defendants' answers to questions on arrests of their friends, the consumption of illegal drugs by their friends, or whether their friends and family had been victims of crime, raises severe doubts regarding the relevance of all these questions to the purposes of processing.[226]

3.4.3 Principle of accuracy and AI

With regard to the principle of accuracy, the GDPR does not differentiate between data provided by data subjects and observed, derived, or inferred ones. According to the WP29, the principle should be observed by controllers throughout processing, especially when personal data are collected and analysed to construct profiles, which will be afterwards applied to take impactful decisions about individuals.[227] If data are incorrect or outdated, profiles and decisions based on them will be faulty, leading to misrepresentations of data subjects' health, credit, or insurance risk.[228] Meanwhile, the accuracy of individual data does not guarantee the accuracy of the entire data-set, which could be misrepresentative or skewed by bias.[229] Therefore, the WP29 calls data controllers to implement appropriate measures so that the data being processed are steadily accurate and up to date.[230] In this effort, it is important to clearly inform data subjects about processing, in order for them to fix any mistakes in advance through their right to rectification (Article 13 (2) b)) and thereby restore the quality of their data.

---

[225] Information Commissioner's Office, 'Big Data, Artificial Intelligence, Machine Learning and Data Protection'.
[226] Julia Angwin, 'Sample COMPAS Risk Assessment', accessed 19 November 2018, https://www.propublica.org/documents/item/2702103-Sample-Risk-Assessment-COMPAS-CORE.
[227] Article 29 Data Protection Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679'.
[228] Article 29 Data Protection Working Party.
[229] Article 29 Data Protection Working Party.
[230] Article 29 Data Protection Working Party.

However, ethically problematic decisions reached by AI are often attributed to biased training data. AI development overlaps with Big Data, whose messiness is regarded as advantageous. When only limited, small data were available, it was imperative that at least those data gathered were of high-quality and accuracy or, shortly put, *'[t]he obsession with exactness is an artifact of the information-deprived analog era.'*[231] These times have irrevocably passed and in our digitally connected world lowering the bar of precision enables larger collections of data at reduced costs.[232] Thus, instead of worrying about the exactitude of individual data, Mayer-Schönberger and Cukier encourage embracing their messiness.[233] At the same time, Big Data analysis through ML rests heavily on correlations, whose validity is unreasonably presumed. Consequently, tensions arise between the principle of accuracy and data-intensive AI.

Indeed, when Loomis, one of the defendants aggrieved by COMPAS, challenged its use in court, the right to be sentenced based on accurate information was one of the legal bases upon which he established his claims.[234] Although the Wisconsin Supreme Court eventually did not rule in his favour, it recognised the need to constantly monitor the accuracy of such algorithms.[235] Dressel and Farid's research sharply demonstrated that predictions reached by COMPAS are as accurate as those reached by a random group of inexpert volunteers.[236] Similarly, the data processing which resulted in the classification of black people as gorillas by Google Photos was in blatant violation of the principle of accuracy.

From the soft law standpoint, in its Briefing *Understanding artificial intelligence*, the EPRS mentioned the quality of data as one of the key issues that must be warranted in order to avoid biases in AI.[237] The European Political Strategy Centre (EPSC) in its Strategic Note *The Age of Artificial Intelligence* observed that a lack of diversity and interdisciplinarity in AI development spoils the accuracy of data by instilling biases in them.[238] In its Opinion *Artificial intelligence - The consequences of artificial intelligence on the (digital) single market,*

[231] Mayer-Schönberger and Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, 40.
[232] Mayer-Schönberger and Cukier, 40.
[233] Mayer-Schönberger and Cukier, 40.
[234] 'State v. Loomis', Harvard Law Review, accessed 12 November 2018, https://harvardlawreview.org/2017/03/state-v-loomis/.
[235] 'State v. Loomis'.
[236] Julia Dressel and Hany Farid, 'The Accuracy, Fairness, and Limits of Predicting Recidivism', *Science Advances* 4, no. 1 (January 2018): eaao5580, https://doi.org/10.1126/sciadv.aao5580.
[237] Reillon, 'Understanding Artificial Intelligence'.
[238] European Political Strategy Centre, 'The Age of Artificial Intelligence: Towards a European Strategy for Human-Centric Machines', EPSC Strategic Notes (European Commission, 27 March 2018), https://ec.europa.eu/epsc/sites/epsc/files/epsc_strategicnote_ai.pdf.

production, consumption, employment and society, the EESC skillfully explained why this is the case:

> *The development of AI is currently taking place within a homogenous environment principally consisting of young, white men, with the result that (whether intentionally or unintentionally) cultural and gender disparities are being embedded in AI, among other things because AI systems learn from training data. This data should be accurate and of good quality, diverse, sufficiently detailed and unbiased. There is a general tendency to believe that data is by definition objective; however this is a misconception. Data is easy to manipulate, may be biased, may reflect cultural, gender and other prejudices and preferences and may contain errors.*[239]

As the propagation of undiversified AI ecosystems is a roadblock to accurate and unbiased AI, the EPSC prescribes, on one side, the allocation of funding to create incentives for talent diverse in terms of disciplines, gender, or ethnicity to join the field and, on the other side, the collaboration among researchers with economic, social, historical, ethical, and anthropologic expertise to assess AI technologies.[240]

An indirect solution advanced by the Commission is the free flow of high-quality and accurate non-personal public data, so that these are used in lieu to train AI.[241] To that end, a Regulation on a framework for the free flow of non-personal data in the EU is on the cusp of coming into force, while the Public Sector Information Directive is under review to encourage re-use of public sector data for the development of AI and other emerging technologies.[242] Conversely, a proposed Directive on copyright, which allows data mining of copyrighted content only for scientific research, will slow down commercial AI development.[243] Indeed, Levendowski has argued that a major obstacle to unbiased data for AI is copyright and intellectual property law, which restricts access to quality, accurate data.[244] Fortunately, the

---

[239] Muller, 'Artificial Intelligence – The Consequences of Artificial Intelligence on the (Digital) Single Market, Production, Consumption, Employment and Society'.
[240] European Political Strategy Centre, 'The Age of Artificial Intelligence: Towards a European Strategy for Human-Centric Machines'.
[241] Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs, 'On the Road to Automated Mobility: An EU Strategy for Mobility of the Future'; Directorate-General for Communications Networks, Content and Technology, 'Artificial Intelligence for Europe'.
[242] 'Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a Framework for the Free Flow of Non-Personal Data in the European Union', Pub. L. No. 32018R1807, OJ L 303 (2018), http://data.europa.eu/eli/reg/2018/1807/oj/eng; Mar Negreiro, 'Re-Use of Public Sector Information', Briefing (European Parliamentary Research Service, European Parliament, November 2018), http://www.europarl.europa.eu/RegData/etudes/BRIE/2018/628312/EPRS_BRI(2018)628312_EN.pdf.
[243] Tambiama Madiega, 'Copyright in the Digital Single Market', Briefing (European Parliamentary Research Service, European Parliament, July 2018), http://www.europarl.europa.eu/RegData/etudes/BRIE/2016/593564/EPRS_BRI(2016)593564_EN.pdf.
[244] Amanda Levendowski, 'How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem', *Washington Law Review* 93 (24 July 2017): 579–630.

Commission announced the creation of European data spaces with robust and GDPR compliant data-sets available for AI development.[245]

3.4.4 Principle of storage limitation and AI

Obstacles exist in aligning AI with the principle of storage limitation, which demands that data controllers save personal data solely for limited periods. This principle is a facet of data minimisation and is interlaced with the subjects' right to erasure (right to be forgotten).

By design, AI processes and accordingly stores troves of data, as the time and financial resources needed for storage are declining so much that it is nowadays less costly to store than discard data.[246] As AI finds new correlations even among the same data-set, retaining data over time allows for their future exploitation. By restricting the retention of data, AI systems lose a considerable part of their pool of accumulated data used for continuous training and testing, which hinders their improvement and, consequently, their compliance with the principle of accuracy. Thus, AI is in clash with storage limitation as a GDPR principle and as a good practice in records management. A corollary result might be that being forced to dispose of data within specific time limits will lessen the processing of historical data by AI, thereby curtailing episodes of historical bias.

3.4.5 Conclusion

In theory, the principles of purpose limitation, data minimisation, accuracy, and storage limitation could be helpful against biased data, as the European Parliament also observed in Article 20 of its Resolution.[247] Nevertheless, compliance of AI with these principles is, in essence, at odds with its basic features. This implies that either AI systems must be designed in fundamentally different ways or legal provisions should explicitly address these blind spots.

3.5 Data Protection Impact Assessments under Article 35 GDPR

The EU's legal arsenal includes another noteworthy tool applicable to AI, the Data Protection Impact Assessment (DPIA). Amid the data controllers' general obligations under the principle

[245] Directorate-General for Communications Networks, Content and Technology, 'Coordinated Plan on Artificial Intelligence', 7 December 2018.
[246] Information Commissioner's Office, 'Big Data, Artificial Intelligence, Machine Learning and Data Protection'.
[247] European Parliament et al., 'Civil Law Rules on Robotics'.

of accountability, Article 24 (1) GDPR notes that, while establishing measures in compliance with the Regulation, they should gauge the likelihood and severity of risks that processing may pose to natural persons' rights and freedoms. Viewed under this overarching obligation, Article 35 designates data controllers' obligation to conduct a DPIA. DPIAs concerning crime-related data are provided for in Article 27 of the Police Directive.

Pursuant to Article 35 (1) GDPR, types of processing, especially those employing new technologies, which are *'likely to result in a high risk to the rights and freedoms of natural persons'* should be preceded by *'an assessment of the impact of the envisaged processing operations on the protection of personal data'*. In paragraph 3, the same Article provides a list of three cases where DPIAs are mandated. The first case refers to systematic and wide-ranging automated processing of personal data, including profiling, which leads to decisions with legal or significantly similar implications for natural persons, whereas the second case encompasses large-scale processing of the special categories of data laid down in Article 9 (1) or the crime-related data in Article 10. The third case refers to systematic, large-scale monitoring of publicly accessible areas. Given the indicative character of this list, there might be high-risk cases outside these cases but still subject to DPIAs. If there are doubts on whether the conditions are met, data controllers are advised to conduct them anyway as good practice confirming compliance with the GDPR.[248] If the conditions required for a DPIA are certainly not met, the controllers' aforesaid general obligation to take precautions against risks to individuals' rights and freedoms is still binding.

Although the GDPR does not explicitly define DPIAs, their minimum content is deduced from Article 35 (7). Within their context, data controllers describe the envisaged operations and purposes of processing and evaluate whether the former are necessary and proportional to the latter. Moreover, they ought to assess the risks to data subjects' rights and freedoms as well as expound the means they have determined to address them, protect personal data and, in general, comply with the GDPR. According to the WP29, subject to risk by processing are primarily the rights to data protection and privacy and peripherally fundamental rights such as the right to liberty and freedom of thought, conscience, religion, speech, or movement.[249] This active consideration of possible risks to determine the applicability of the provision, without sidelining individuals' protection in the form of established rights, is

[248] Article 29 Data Protection Working Party, 'Guidelines on Data Protection Impact Assessment (DPIA) and Determining Whether Processing Is "Likely to Result in a High Risk" for the Purposes of Regulation 2016/679', 4 October 2017, http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611236.
[249] Article 29 Data Protection Working Party.

representative of the 'risk-based approach' firstly adopted in the Data Protection Directive and followed by the GDPR.[250]

During a DPIA, controllers examine a single data processing operation or a set of operations posing similar risks (Article 35 (1)). As Recital 92 explains, the latter is preferable for practical or financial reasons, for instance when the same application or processing platform is about to be deployed by multiple public authorities or data controllers of the same industry. These provisions advance a systematic exploration of the built-in risks of new situations, especially when new technologies of data processing emerge. This means that if the risks of an application or processing platform have been evaluated in similar contexts under a previous DPIA, they will not need to be anew examined. Also, the WP29 suggests that providers conduct DPIAs for technological products, e.g. a piece of hardware or software, which will be supportive to the different DPIAs subsequently conducted by controllers based on the specific ways and contexts in which they will deploy the same technological product.[251]

Under Article 36, should the results of a DPIA reveal a high risk of processing, which cannot be mitigated with measures by the data controller, the latter must preliminarily consult with the data protection authority.

3.5.1 DPIAs and AI

The WP29 explicates nine criteria for evaluating whether processing is likely to be high-risk. Among these, AI-enabled decision-making, especially in the case of COMPAS and iBorderCtrl, easily falls within the remit of one or more of the following:

- Evaluation or scoring, especially based on data subjects' work performance, finances, health, personal preferences, reliability, or location (Recitals 71 and 91).[252]
- Automated decision-making with legal or similarly significant effects (Article 35 (3) a)).[253]
- Processing of special categories of data of Article 9 or crime-related data of Article 10, in which cases the private or public character of data is of relevance.[254]

[250] Article 29 Data Protection Working Party, 'Statement on the Role of a Risk-Based Approach in Data Protection Legal Frameworks', 30 May 2014, 29, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp218_en.pdf; 'Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data', Pub. L. No. 31995L0046, OJ L 281 (1995), http://data.europa.eu/eli/dir/1995/46/oj/eng.
[251] Article 29 Data Protection Working Party, 'Guidelines on Data Protection Impact Assessment (DPIA) and Determining Whether Processing Is "Likely to Result in a High Risk" for the Purposes of Regulation 2016/679'.
[252] Article 29 Data Protection Working Party.
[253] Article 29 Data Protection Working Party.
[254] Article 29 Data Protection Working Party.

- Large-scale processing, which the WP29 interprets as depending on the number of data subjects concerned; the volume and range of data; the duration and geographical extent of processing.[255]
- Matching or combining data-sets beyond the reasonable expectations of data subjects.[256]
- Data concerning vulnerable data subjects (Recital 75), such as children, employees, segments of the population requiring special protection, and generally cases of power imbalance between data subjects and controllers.[257]
- Innovative use of existing technology or application of new technology, compared to the achieved state of technological knowledge (Recitals 89 and 91), due to the unknown personal and social consequences of such cases.[258]
- Processing which allows, modifies or refuses data subjects' access to exercising a right, using a service, or entering a contract (Article 22, Recital 91).[259]

Similarly, in the list published by ICO, AI is prominently mentioned amidst types of processing requiring a DPIA and is included as indicative of all three mandatory cases of DPIAs, as these are mentioned in Article 35 (3).[260] Furthermore, ICO underlines that any automated processing with AI or ML is likely to require a DPIA, even if it includes human interjection.[261]

However, as AI entails threats of discriminatory bias against individuals, the European Union Agency for Fundamental Rights (FRA) notes that such impact assessments should be broader and geared towards risks for biases and discrimination, even proxy discrimination, in the automated decision-making process and its outputs.[262] Mantelero finds that, in the case of AI and Big Data, DPIAs are insufficient in including ethical and societal considerations, which extend beyond the individual dimension of fundamental rights and have collective consequences, such as prejudices and discrimination against social groups.[263] As an alternative to the mainly focused on data security DPIAs, he articulates a rights-based and values-oriented model of such assessments in the form of a voluntary Human Rights, Ethical and Social Impact Assessment (HRESIA), comprised of a self-assessment tool and an ad hoc expert committee.[264]

---

[255] Article 29 Data Protection Working Party.
[256] Article 29 Data Protection Working Party.
[257] Article 29 Data Protection Working Party.
[258] Article 29 Data Protection Working Party.
[259] Article 29 Data Protection Working Party.
[260] Information Commissioner's Office, 'Data Protection Impact Assessments (DPIAs)', 14 May 2018, https://ico.org.uk/media/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/data-protection-impact-assessments-dpias-1-0.pdf.
[261] Information Commissioner's Office.
[262] European Union Agency for Fundamental Rights, '#BigData: Discrimination in Data-Supported Decision Making' (Vienna, 29 May 2018), http://fra.europa.eu/en/publication/2018/big-data-discrimination.
[263] Alessandro Mantelero, 'AI and Big Data: A Blueprint for a Human Rights, Social and Ethical Impact Assessment', *Computer Law & Security Review* 34, no. 4 (1 August 2018): 754–72, https://doi.org/10.1016/j.clsr.2018.05.017.
[264] Mantelero.

Relatedly, the AI Now Institute of New York University draws on DPIAs and acknowledges their importance to suggest a more targeted tool, Algorithmic Impact Assessments (AIAs). Compared to DPIAs, their suggested assessments publically engage the communities likely to be affected by the AI systems under deliberation and specifically evaluate the incorporation of AI by public authorities.[265]

Nonetheless, even if the scope of DPIAs changes to accommodate ethical concerns of AI algorithms and Big Data, they are still susceptible to ritualism and creative compliance.[266] It is to be feared that expectations of comprehensive and meaningful risk assessments will be nullified by a perfunctory observance of the rules, detached from their rationale or spirit. This check-box mentality, falsely reassuring data controllers that just because they have routinely completed these assessments they need not be vigilant or accountable any more, should by any means be counteracted.[267]

3.5.2 Conclusion

Summarising, either in the form of DPIAs, AIAs, HRESIAs, or any other AI-related assessments, estimating the potential hazards and impact of AI systems should be reconceptualised so as to focus on the identification of threats to fairness and equality. If this happens and also DPIAs are not approached as another box to be ticked, they could halt discriminatory decisions before these are made and data subjects are in effect hurt.

3.6 Automated decision-making under Article 22 GDPR

Article 22 features prominently in all AI-targeted soft law instruments. In paragraph 1, it provides for data subject's *'right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.'* Autonomy, honour, and human dignity lie at the heart of this right, inasmuch as treating individuals as fellow persons instead of mechanically calculated percentages and allowing them to have their personally important matters evaluated by other humans are manifestations of respect to these protected notions.[268] Contrarily, the

[265] Dillon Reisman et al., 'Algorithmic Impact Assessment: A Practical Framework for Public Agency Accountability' (AI Now, April 2018), https://ainowinstitute.org/aiareport2018.pdf.
[266] Reuben Binns, 'Data Protection Impact Assessments: A Meta-Regulatory Approach', *International Data Privacy Law* 7, no. 1 (February 2017): 22–35, https://doi.org/10.1093/idpl/ipw027.
[267] Reisman et al., 'Algorithmic Impact Assessment: A Practical Framework for Public Agency Accountability'.
[268] Tal Zarsky, 'Transparent Predictions', *University of Illinois Law Review* 2013, no. 4 (10 September 2013): 1503–70.

'chilling effects' produced when humans sense that they are being judged and circumscribed by impersonal machines oppose their self-determination and freedom of expression.[269]

3.6.1 Scope of application

Article 22 updates Article 15 of the Data Protection Directive and its provisions are similarly included in Article 11 of the Police Directive. For a much-needed exegesis of the Article, the WP29 issued guidelines on automated decision-making, which they found increasingly used in both private and public sector.[270]

Of crucial interest to the WP29 is the deployment of algorithms for classification based on profiling methods, as it preserves and exacerbates stereotypes and social segregation.[271] By classifying individuals into rigid categories, it curbs their freedom to choose among various goods and construct their own identity. Errors and biases in data or the automated process lead to incorrect classifications and thereby inaccurate predictions.[272] In turn, inaccurate predictions are not mere statistical problems; they result in judgements deleterious to individuals, such as unjustified denial of resources or opportunities and discrimination.[273] Similar dismay against inferential decisions reached by algorithms was conveyed by the WP29 in 2013.[274] In view of these critiques, the WP29 echoes to a large extent the epistemic and normative types of ethical concerns discussed in Chapter 2.

The WP29 interprets *'solely automated'* decision-making as occurring exclusively through technological means, absent human involvement. Nonetheless, as minimal human involvement could be fabricated to elude the scope of these provisions, this requirement is significantly nuanced. If human agents participate in the process but do not, in effect, influence its outputs, no human involvement is affirmed.[275] Contrarily, if human agents exert meaningful oversight of the outputs and are equipped with the authority to change them, human involvement is asserted and decision-making is not solely automated.[276] Similarly, the main

[269] Sandra Wachter and Brent Mittelstadt, 'A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI', *Columbia Business Law Review* Forthcoming (13 September 2018), https://papers.ssrn.com/abstract=3248829.
[270] Article 29 Data Protection Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679'.
[271] Article 29 Data Protection Working Party.
[272] Article 29 Data Protection Working Party.
[273] Article 29 Data Protection Working Party.
[274] Article 29 Data Protection Working Party, 'Opinion 03/2013 on Purpose Limitation'.
[275] Article 29 Data Protection Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679'.
[276] Article 29 Data Protection Working Party.

justification of the ruling against Loomis in State v. Loomis was that decisions reached by COMPAS were accompanied by officials' evaluations of defendants' recidivism.[277] Thus, in GDPR parlance, data processing was not solely automated.

The automated process refers to a decision concerning an 'individual' data subject, as suggested by the headings of Article 22 and Section 4, without enunciating whether it must be a final or just an interim decision. Recital 71 seems to include intermediate steps of the decision-making process in Article 22, as it refers to *'the right not to be subject to a decision, which may include a measure'*. Having said that, of all possible automated decisions, the Article applies only to those which incur legal or similarly significant effects. The effects can be material or immaterial, e.g. affecting the subject's dignity, integrity, or reputation, and must be negative for the individual.[278] Furthermore, WP29's position is that these effects might derive from processing other people's data, meaning a group of which the data subject is assumed to be a member because of shared characteristics, as is the case when one's credit card limit is curtailed or extended based on analyses of nearby customers' transactions.[279] As Article 22 only stipulates that the decision applies to an individual, regardless of whether the automated process is based on information about members of a whole group, it is applicable to such cases as well.[280]

The GDPR does not delimit legal or similarly significant effects. For Kamarinou et al., decisions with legal effects equate to binding decisions or decisions from which legal obligations emanate.[281] For the WP29, they allude to effects on one's legal status or rights, with examples including the decision to grant a social benefit to someone, refuse entry at the border at someone, subject one to surveillance, and terminate one's mobile phone service.[282] For instance, COMPAS and iBorderCtrl would fall under the scope of decisions with legal effects.

Even more puzzling are the effects which are similarly significant to legal ones. The ICO contends that the application of ML in Big Data analysis, when used to profile individuals

[277] 'State v. Loomis'.
[278] Dimitra Kamarinou, Christopher Millard, and Jatinder Singh, 'Machine Learning with Personal Data', Legal Studies Research Paper 247/2016 (Queen Mary University of London, School of Law, 7 November 2016), https://ssrn.com/abstract=2865811; 'Rights Related to Automated Decision Making Including Profiling', 6 August 2018, https://icoumbraco.azurewebsites.net/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/rights-related-to-automated-decision-making-including-profiling/.
[279] Article 29 Data Protection Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679'.
[280] Kamarinou, Millard, and Singh, 'Machine Learning with Personal Data'.
[281] Kamarinou, Millard, and Singh.
[282] Article 29 Data Protection Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679'.

and extract decisions about them, significantly affects them in that sense.[283] Indeed, Recital 71 states as examples the automatic refusal of an online credit application and automated recruiting practices, thus Amazon's recruitment AI would be among these. For the WP29, this implies that the decision impacts data subjects' circumstances, behaviour, or choices to a significant extent, with the worst outcome being their exclusion or discrimination.[284]

In any case, proving which decision-making processes fall under Article 22 and which not imposes an undue burden to data subjects, whose protection is conditional upon a vague threshold of requirements.

3.6.2 Prohibition of automated decision-making and derogations

In the literature, it is debated whether Article 22 (1) establishes a right to objection or a prohibition.[285] The difference between these two interpretations is that the former places the onus on individuals to find out the existence of automated processing and timely object to it. The WP29 supports the latter interpretation, which means that solely automated decision-making is generally prohibited and individuals have the right not to be subject to a decision based solely on automated decision-making, unless one of the following three derogations carved out by Article 22 (2) applies.[286] Hence, data subjects are by default protected against solely automated decision-making.

The first derogation wherein solely automated decision-making is allowed is its utilisation for entering or performing a contract. This might happen when automated decision-making has the potential to ensure more consistent, less erroneous conclusions or when the quantity of data to be processed exceeds human capacities. In line with Recital 71, the WP29 exhorts a narrow interpretation of the exception in the sense that automated decision-making must be necessary for the contract, not merely useful.[287] If other, less intrusive means are available, then it is not necessary and these other means should be preferred in its place. The second derogation refers to the authorisation of automated decision-making by EU or Member State law. According to Recital 71, this is for example applicable when it is used for monitoring

[283] Information Commissioner's Office, 'Big Data, Artificial Intelligence, Machine Learning and Data Protection'.
[284] Article 29 Data Protection Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679'.
[285] Wachter, Mittelstadt, and Floridi, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation'.
[286] Article 29 Data Protection Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679'.
[287] Article 29 Data Protection Working Party.

or preventing fraud and tax evasion or for warranting the security of a service offered by the data controller. The third derogation is when data subjects have given explicit consent to automated decision-making. The WP29 postulates that this exception is satisfied by an express statement, not just any affirmation.[288]

Even when these derogations permit solely automated decision-making, Article 22 (3) establishes an additional layer of protection for data subjects, meaning that their rights and legitimate interests need to be safeguarded by data controllers. Appropriate safeguards indicatively include the rights to insist on human intervention, express their point of view, and challenge the decision.

As regards the right to obtain human intervention, in order for processing to bypass the scope of being solely automated, the intervention should be performed by someone with competency to review and change the decision. On the flip side, the fact remains that human intervention decelerates the decision-making process, whilst rapid turnaround is one of the main appeals that the introduction of AI holds in the first place. Additionally, due to the pervasive air of authority of AI, human reviewers will be disinclined to challenge and meaningfully re-examine its de facto valid conclusions. Similarly, when it comes to expressing their viewpoint, whose conclusions are more likely to be convincing: that of a subjective, directly involved in the case individual or that of an impartial, mathematical model? The answer is left to the readers' wisdom. Insofar as the right to challenge the decision is concerned, it is dubious whom should data subjects appeal for a hearing or review after a decision has been made. Apropos that, Kamarinou et al. uphold that, instead of a human reviewer, data subjects could appeal to an AI system to review the previous automated decision because of its potentially high level of error resistance.[289] Yet, this solution is fatally prone to the same ethical and legal shortcomings of AI decision-making portrayed so far.

3.6.3 A right to explanation and unexplainable AI

The utmost bone of contention is whether Article 22 grants data subjects with a 'right to explanation' and, if yes, what it entails. If the GDPR indeed establishes a right to explanation, it burdens individuals or companies processing EU citizens' data with the obligation to provide them with meaningful explanations of how their automated systems reach decisions. Wachter et al. put forward two possible kinds of explanation in automated decision-making: on one side,

[288] Article 29 Data Protection Working Party.
[289] Kamarinou, Millard, and Singh, 'Machine Learning with Personal Data'.

explaining system functionality of an automated decision-making system means construing its logic, significance, anticipated consequences, and general functionality; on the other side, unravelling specific decisions involves construing their rationale, reasons, and individual circumstances.[290] Explanations of system functionality are ex ante, i.e. prior to the automated decision-making process, or ex post, i.e. after the automated decision-making process, whereas explanations of specific decisions are solely given ex post.[291]

Arguments granting data subjects with a right to explanation stem mainly from Recital 71. Among the suitable safeguards to be ensured in automated decision-making, Recital 71 enumerates the ones contained in 22 (3), i.e. the data subjects' right to information, to express their viewpoint, and to contest the decision, with the further addition of the right *'to obtain an explanation of the decision reached after such assessment'*. This implies the right to an ex post explanation of specific decisions according to the aforesaid typology of Wachter et al. Among adherents to this interpretation are the UK House of Commons, the ICO, and the FRA, as explicitly mentioned in their related reports.[292]

Opposing interpretations rely on the fact that, despite their uncontested significance in EU law, Recitals are not legally binding. Deprived of the ability to raise legitimate expectations or create rules on their own, their role is to proffer interpretative guidance in cases of ambiguity, but no such ambiguity exists in the wording of Article 22 (3) regarding the minimum requirements with which data controllers ought to comply.[293] Besides, the fact that the right to explanation is only included in a Recital plausibly reflects a deliberate legislative choice.[294] In support of this argument, Wachter et al. propound a historical interpretation of the controversial Article and elucidate that the European Parliament's recommendation to include a *'right to obtain human assessment and an explanation of the decision reached after such assessment'* in the main body of the GDPR was followed by the European Council's suggestion to move this right to the Recitals, a suggestion which ultimately prevailed in the texts adopted.[295]

[290] Wachter, Mittelstadt, and Floridi, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation'.
[291] Wachter, Mittelstadt, and Floridi.
[292] Science and Technology Committee, 'Robotics and Artificial Intelligence' (UK: House of Commons, 10 December 2016), https://publications.parliament.uk/pa/cm201617/cmselect/cmsctech/896/896.pdf; Information Commissioner's Office, 'Big Data, Artificial Intelligence, Machine Learning and Data Protection'; European Union Agency for Fundamental Rights, '#BigData: Discrimination in Data-Supported Decision Making'.
[293] Wachter, Mittelstadt, and Floridi, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation'.
[294] Wachter, Mittelstadt, and Floridi.
[295] Wachter, Mittelstadt, and Floridi.

Therefore, this omission was caused intentionally after weighing the competing versions, not because of negligence on behalf of the legislative body, so it should be respected.

The amalgamation of Articles 13 (2) f) and 14 (2) g), which apply respectively when data are obtained directly from data subjects or third parties, is additionally interpreted by Goodman and Flaxman as engendering a right to explanation.[296] Both Articles assert that data subjects are entitled to be aware of the application of automated decision-making, as this is prescribed in Article 22, and receive information about the logic involved, the significance and anticipated consequences of processing. Such information must be meaningful to them, which denotes that communicating random facts or highly technical particulars about the system is insufficient. In any case, the sufficiency of information, according to Recital 60, is assessed considering the context of processing. Contra Goodman and Flaxman, Wachter et al. refute the derivation of a right to an ex post explanation of specific decisions from Articles 13 and 14. As the notification duties of these Articles have to be fulfilled before decision-making takes place, they could only grant a right to an ex ante explanation.[297] However, following their typology, ex ante explanations cover solely system functionality, not specific decisions. Hence, Articles 13 and 14 cannot establish a right to an ex post explanation of specific decisions.[298]

Nonetheless, Wachter et al. leave open the possibility for deriving a right to explanation from the right to contest automated decisions (Article 22 (3)) and its accompanying right to fair trial and effective remedy (Article 47 CFREU).[299] Court proceedings involve examination of witness statements, testimonies, documents and other sources of evidence in support of a decision. Absent an explanation of the rules and models upon which an automated decision was reached, it will be troublesome, even unfeasible for data subjects to challenge it in court and for the judiciary to effectively review relevant evidence and claims.[300] As Justice Abrahamson conceded referring to the State v. Loomis adjudication, the *'lack of understanding of COMPAS was a significant problem'* and *'the court needed all the help it could get'*.[301] Given the crucial role of a right to fair trial for democracy itself, the debate about a right to explanation appears heavily loaded.

---

[296] Goodman and Flaxman, 'European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"'.
[297] Wachter, Mittelstadt, and Floridi, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation'.
[298] Wachter, Mittelstadt, and Floridi.
[299] Wachter, Mittelstadt, and Floridi.
[300] Wachter, Mittelstadt, and Floridi; Citron, 'Technological Due Process'; John Zerilli et al., 'Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?', *Philosophy & Technology*, 5 September 2018, 1–23, https://doi.org/10.1007/s13347-018-0330-6.
[301] 'State v. Loomis'.

Towards settling this dispute, the WP29 holds that data controllers' obligation is confined to explaining in simple terms the rationale behind or the criteria employed in reaching a decision, but it is not necessary to offer a complex explanation or full disclosure of the AI algorithms.[302] This stance is deemed pragmatic, since the growth and complexity of ML models make it tough for controllers to understand how they work or, a fortiori, explain it to data subjects with fluctuating degrees of technical understanding. From the soft law outlook, the EPRS adopts the same moderate interpretation and submits that Article 22 will be challenging for AI developers and for successfully eliminating biases, as Narrow AI decisions are currently indecipherable by users and developers.[303]

At the other end of the scale lies the detrimental impact that AI-enabled decisions bear on individuals, e.g. being denied employment by Amazon's recruitment AI, being turned away at a border crossing because of iBorderCtrl, and being sentenced to incarceration by COMPAS. Due to such cases, Mittelstadt et al. entertain the thought of limiting the use of automation systems in critical contexts if their rationale is obfuscated.[304] For instance, the US Fair Credit Reporting Act precludes AI systems from credit scoring, as they cannot comply with consumers' rights to know on demand the reasons underlying negative decisions.[305] In the same direction, the EESC recommends the segmentation of decision-making procedures to these suitable or not for delegation to AI alongside a division of cases wherein human intervention is desirable or mandatory.[306] Moving this line of thought further, Ananny and Crawford support that when the complexity of a system is such that even individuals with a total overview of it cannot articulate its failed or successful reasoning, it is doubted whether the system should be developed at all.[307]

Martin adds that, by not holding AI development firms accountable on the basis of the inexplicability of AI, they are instead incentivised to design more inscrutable algorithms in order to escape their obligations for information and explanation.[308] As seen in Chapter 2, the inscrutability of AI may be used as a pretext for sub rosa discrimination. To prevent such cases, the WP29 clarifies that complexity should not be invoked as an excuse to override data

[302] Article 29 Data Protection Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679'.
[303] Reillon, 'Understanding Artificial Intelligence'.
[304] Mittelstadt et al., 'The Ethics of Algorithms'.
[305] Burrell, 'How the Machine "Thinks"'.
[306] Muller, 'Artificial Intelligence – The Consequences of Artificial Intelligence on the (Digital) Single Market, Production, Consumption, Employment and Society'.
[307] Binns, 'Fairness in Machine Learning'.
[308] Martin, 'Ethical Implications and Accountability of Algorithms'.

subjects' rights to information.[309] Furthermore, as higher degrees of complexity and inexplicability of an AI system limit the role of other human agents in decision-making, AI developers should be held accountable for decisions taken by these systems.[310] The voluntary development of an AI algorithm which is known to function in value-laden, intricate, and hard to understand ways and its equally deliberate sale to be used in decision-making contexts is sufficient to render AI developers part of the decision system, bound by its concomitant responsibilities.[311]

Given that at its current phase AI does not abide by the aforementioned requirements of explicability, the EPRS, the EESC and the European Parliament underline that it is crucial for automated decision-making systems to be developed in a manner that guarantees their interpretability and monitorability by humans.[312] To achieve this, one of the Commission's policy suggestions is to support scientific research into the explainability of AI (Article 3 (3)).[313] At the same time, the European Parliament asks for all the transactions in which an AI robot participates along with the logic guiding its decisions to be constantly recorded in its software (Article 12).[314] According to the EESC, the inner workings and decisions of AI should be accessible to human understanding not only at the moment of their emergence but also retrospectively.[315] On the other side, a type of 'counterfactual explanations' is suggested by scholars to alternatively support data subjects' future actions without necessarily demystifying the internal mechanisms of AI.[316]

Overall, without clear-cut ways to satisfy these requirements in the near future, the acceptance, sustainable development, and application of AI is at risk. Eventually, this debate depicts the value-laden character of AI development. As the EGE captured in its Statement, when thinking about the explainability of AI, we need to ask:

> *'Which values do these systems effectively and demonstrably serve? Which values underpin how we design our policies and machines? Around which*

[309] Article 29 Data Protection Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679'.
[310] Martin, 'Ethical Implications and Accountability of Algorithms'.
[311] Martin.
[312] Reillon, 'Understanding Artificial Intelligence'; Muller, 'Artificial Intelligence – The Consequences of Artificial Intelligence on the (Digital) Single Market, Production, Consumption, Employment and Society'; European Parliament et al., 'Civil Law Rules on Robotics'.
[313] Directorate-General for Communications Networks, Content and Technology, 'Artificial Intelligence for Europe'.
[314] European Parliament et al., 'Civil Law Rules on Robotics'.
[315] Muller, 'Artificial Intelligence – The Consequences of Artificial Intelligence on the (Digital) Single Market, Production, Consumption, Employment and Society'.
[316] Sandra Wachter, Brent Mittelstadt, and Chris Russell, 'Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR', *Harvard Journal of Law & Technology* 31, no. 2 (2 November 2017), http://arxiv.org/abs/1711.00399.

> *values do we want to organise our societies? And which values are we letting to be undermined –openly or silently- in the technological progress and utility trade-off?'*[317]

3.6.4 Additional rights of data subjects

If the relevant provisions are sufficiently unsnarled regarding their application to AI, the qualified prohibition of automated decision-making in combination with rights to obtain information, to have recourse to human review, and to challenge automated decisions could be used as a lever against automated discrimination.

Outside the scope of Article 22 or as supplements thereof, data subjects have a panoply of updated rights in their legal armoury: the right of access to their personal data (Article 15 (1) h)); the right to rectification (Article 16), which applies to personal data used as input and those created as output, i.e. a data subject's profile or score; the right to erasure (Article 17); the right to restrict processing (Article 18), which should be respected at all stages of data processing; the right to data portability (Article 20); the right to object to processing (Article 21). For example, job applicants could invoke these provisions if evaluated by Amazon's recruitment AI. However, the volume and scale of data processed by AI hinder the isolation and separate treatment of data belonging to a specific individual, which might thwart a meaningful realisation of these rights. Relatedly, their actualisation is likely to affect other data subjects' rights. If an individual decides to correct or erase their data from the data-set, this will modify the overall ML model built upon this data-set, leading to different conclusions for the remaining data subjects. Therefore, exercising such rights is not all plain sailing for data subjects nor AI developers.

3.7 Lawful, fair, and transparent processing under Article 5 (1)

In the GDPR, Article 5 (1) a) establishes the principles of lawfulness, fairness, and transparency. As the black box problem of AI similarly hampers their implementation, Article 22 is commonly probed in tandem or merged with the principle of transparency, so they are examined here successively. Notably, the Police Directive provides solely for lawful and fair processing in its equivalent Article 4 (1) a), whilst transparency is mentioned in Recital 26.

[317] European Group on Ethics in Science and New Technologies, 'Artificial Intelligence, Robotics and "Autonomous" Systems'.

3.7.1 Principle of lawfulness and AI

Lawfulness relies on the fulfilment of one of the following six bases of processing under Article 6 (1): consent; performance of contract; compliance with legal obligations; protection of vital interests; public interests; legitimate interests of the controller or a third party. These bases are further limited in solely automated decision-making, for which the general prohibition of Article 22 shall apply as more specific. In all types of AI processing, though, it has been explained that informed consent is hard to be granted on behalf of data subjects as the purposes of processing are unclear at the time of data collection.

Further, Article 9 (1) prohibits the processing of *'data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation'*. In the first instance, controllers must ensure that data from these special categories are not used as input for AI, as this is also clarified in Article 22 (4) on automated decision-making, or, alternatively, they could blinker the AI algorithm to such data. In a second instance, though, even if AI algorithms are not fed special category data, they can infer these from other, non-special category ones, because of their correlative abilities. Through proxy discrimination, seemingly innocent and indifferent data give away sensitive attributes or even re-identify individuals. Such was the case of Amazon's recruitment AI, which used references of the word 'women's' and all-female colleges as proxies for gender and subsequently discriminated against female applicants. To address this issue, the WP29 espouses a broad interpretation of special category data and holds that they fall within the protective scope of GDPR provisions even when they are inferred and not directly collected from data subjects.[318] In search of even stronger protection, Wachter and Mittelstadt further suggest the introduction of a 'right to reasonable inferences'.[319]

In both instances, precluding AI from processing special category data is not unanimously welcomed. Omitting special category data is not only futile, due to their inference from other data, but counterproductive in eliminating discriminatory biases. If their removal is feasible only alongside the non-special category ones from which they are inferred, the AI system will suffer considerable losses in predictive accuracy.[320] By collecting and carefully

[318] Article 29 Data Protection Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679'.
[319] Wachter and Mittelstadt, 'A Right to Reasonable Inferences'.
[320] Barocas and Selbst, 'Big Data's Disparate Impact'.

processing more special category data, rather than less, AI developers can put together more representative and inclusive data-sets, while auditors can better detect discriminatory cases.[321] Imagine a recruitment AI which, in predicting individuals' suitability for a role, takes into account their time needed to complete an online test. If the test is not fully accessible, candidates using assistive technologies to complete it will be disadvantaged.[322] On the contrary, if such information was available for processing, the AI system could be trained on data including candidates with disabilities or its outputs could be a posteriori adjusted to compensate for differences in completion time.[323] Hence, this prohibition does not fit squarely into the demand for unbiased AI.

3.7.2 Principle of fairness and AI

Fairness refers to the effects and expectations of processing on the part of data subjects.[324] As seen in Chapter 2, AI decision-making is prone to yielding discriminatory effects contrary to egalitarian ideals of fairness and equality. Apart from individuals, it is well worth highlighting the effects on social groups, especially underprivileged ones, and society at large, in accordance with the Rawlsian difference principle. COMPAS, for example, does not harm just an individual but a socially salient group, viz. black people. Thus, processing would be fair to the extent that it ensured individual and also group fairness. Moreover, in some cases, the unfair character of such discriminatory effects is visible, e.g. in COMPAS or Amazon's recruitment AI, whereas others at first sight seem frivolous, e.g. in Google Photos, but cause representational harms, which perpetuate biases and should likewise be borne into account. Whether fairness under the GDPR includes such considerations of collective effects and representational harms remains unclear.

Critically, compared to the Data Protection Directive which was silent on cognate issues, Recital 71 of the GDPR, referring to automated decision-making of Article 22, holds that data controllers should ensure that processing does not cause *'discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs,*

[321] Williams, Brooks, and Shmargad, 'How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications', *Journal of Information Policy* 8 (2018): 78–115, https://doi.org/10.5325/jinfopoli.8.2018.0078.
[322] Shari Trewin, 'AI Fairness for People with Disabilities: Point of View' (IBM, 26 November 2018), http://arxiv.org/abs/1811.10670.
[323] Trewin.
[324] Information Commissioner's Office, 'Big Data, Artificial Intelligence, Machine Learning and Data Protection'.

trade union membership, genetic or health status or sexual orientation'. Similarly, Article 11 (3) of the Police Directive forbids discriminatory profiling based on these special categories of data. AI decisions systematically treating a social group differently than others are intuitively considered discriminatory. The EU's anti-discrimination scaffolding, built upon Article 21 CFREU and a skeleton of Directives, prohibits both direct and indirect discrimination.[325] The former refers to a person's treatment in less favourable ways because of membership in a protected social group. The latter, which is most probable in the context of biased AI, obtains when people are disadvantaged by an apparently neutral provision, but is justified when legitimate aims are pursued through appropriate and necessary means.

In both direct and indirect discrimination, proving the existence of differential treatment is incumbent upon the complainants and rarely straightforward. This is worsened when neither complainants nor defendants are aware of the exact variables and processes used by black-boxed AI to generate outputs. The proclaimed objectivity of AI, based upon its statistical backbone and predictive accuracy, makes it easier for defendants to establish that differential treatment is justified. Moreover, it is worth considering that wide-scale automated discrimination may fall outside the personal and material scope of the relevant EU Directives. Except for social groups which have traditionally been targets of marginalisation, AI and Big Data form new, less clear-cut categories based on previously unforeseen attributes, so any biased practices towards them will slip through the cracks of non-discrimination laws.[326] To sum up, it becomes very hard to ascertain the discriminatory effects of AI and the controllers' responsibility to prevent them.

Concerning data subjects' expectations, controllers should evaluate whether the ways in which ML algorithms process data and their results are reasonably anticipated by individuals. As the integration of AI in decision-making becomes seamless, nearly invisible, individuals should be able to form clear expectations about whether decisions affecting them

[325] 'Council Directive 2000/43/EC of 29 June 2000 Implementing the Principle of Equal Treatment between Persons Irrespective of Racial or Ethnic Origin', Pub. L. No. 32000L0043, OJ L 180 (2000), http://data.europa.eu/eli/dir/2000/43/oj/eng; 'Council Directive 2000/78/EC of 27 November 2000 Establishing a General Framework for Equal Treatment in Employment and Occupation', Pub. L. No. 32000L0078, OJ L 303 (2000), http://data.europa.eu/eli/dir/2000/78/oj/eng; 'Council Directive 2004/113/EC of 13 December 2004 Implementing the Principle of Equal Treatment between Men and Women in the Access to and Supply of Goods and Services', Pub. L. No. 32004L0113, OJ L 373 (2004), http://data.europa.eu/eli/dir/2004/113/oj/eng; 'Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the Implementation of the Principle of Equal Opportunities and Equal Treatment of Men and Women in Matters of Employment and Occupation (Recast)', Pub. L. No. 32006L0054, OJ L 204 (2006), http://data.europa.eu/eli/dir/2006/54/oj/eng; Judgment of the Court (Grand Chamber), Case C-236/09, Association belge des Consommateurs Test-Achats ASBL and Others v. Conseil des ministres (European Court of Justice 1 March 2011).

[326] Mantelero, 'AI and Big Data'.

will be formed by a human or artificial agent. More broadly, controllers are bound to exercise data stewardship in a manner that does not deceive or mislead data subjects, thereby inspiring trust.[327]

Overall, effective protection against unfair AI decisions is not guaranteed unless collective and representational harms are also barred, beyond the provisions' individualistic lens. To that end, researchers speculate the use of special category data to build 'algorithmic affirmative actions' for marginalised communities.[328] Additionally, because of their unpredictability by design alongside practical difficulties in establishing indirect discrimination by a supposed impartial but inscrutable algorithm, AI-enabled decisions represent a stumbling block for the principle of fairness.

3.7.3 Principle of transparency and AI

Transparency is not specifically defined in the GDPR, although briefly sketched in Recital 39. The ICO defines transparency in relation to the information offered to data subjects about processing, while the European Parliament refers to it as the possibility to *'supply the rationale behind any decision taken with the aid of AI that can have a substantive impact on one or more persons' lives'*.[329] Thereupon, the debated right to explanation or meaningful information of Article 22 along with Articles 13 and 14 are, amongst others, instantiations of the principle of transparency.

Automated decision-making is largely conducted through opaque processes, which deter individuals from grasping how their data are processed and what kinds of other data are indirectly derived from these.[330] The mechanisms through which AI systems alter their inner constitution are often unknown and unknowable to their developers, resulting in their characterisation as 'black boxes', also adopted by the WP29 to denote the absence of transparency in AI.[331] This opacity of AI diminishes trust on behalf of data subjects and brings about an escalating hesitancy in granting consent for processing, which is particularly worrisome in domains such as healthcare, wherein data sharing could accrue significant

[327] Information Commissioner's Office, 'Big Data, Artificial Intelligence, Machine Learning and Data Protection'.
[328] Anupam Chander, 'The Racist Algorithm?', *Michigan Law Review* 115, no. 6 (2017): 1023–45.
[329] Information Commissioner's Office, 'Big Data, Artificial Intelligence, Machine Learning and Data Protection'; European Parliament et al., 'Civil Law Rules on Robotics'.
[330] Article 29 Data Protection Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679'.
[331] Article 29 Data Protection Working Party.

benefits for humanity.[332] Recital 58 posits that when the level of technological complexity is high and multiple actors are included in the process, it is harder for data subjects to grapple with the conditions of the underway processing, which makes the principle of transparency all the more relevant. Over and above, this applies to automated decision-making in the public sector, as sacrificing the legitimacy of public decisions in the alter of algorithmic opacity would establish a governance model of 'algocracy', as seen in Chapter 2.

To effectuate transparent processing, controllers are encouraged to provide data subjects with the information required under Articles 13 and 14 in a *'concise, transparent, intelligible and easily accessible'* manner (Article 12 (1)). In that sense, mere disclosures of the source code behind the AI system would not be meaningful, as causing information fatigue. The insistence on succinct and lucid supply of information, matched with EU-wide efforts to educate a new generation of technologists, could compensate for the public's limited technical literacy and AI-related awareness, which are partly blamed for epistemic inequalities between individuals and AI developers. The WP29 further suggests that the same AI technologies posing challenges to data protection be leveraged to offer more dynamic and user-centred communication pathways.[333] Indeed, a team of EU/US researchers launched Polisis, an AI delivering visualisations, flow charts, and readable summaries of privacy notices to data subjects.[334]

On the contrary, part of the literature remarks the side effects of transparency. Rendering the details of data processing transparent makes it easier for malevolent individuals to 'game' the AI system, meaning to manipulate it for their benefit.[335] If individuals are aware of the types of data that an AI system recognises and its basic functions, they can deliberately make themselves algorithmically recognisable and feed the AI with data that are more likely to produce favourable to them outputs. Something similar happened when Twitter users abused Tay to produce offensive comments and undermine her operation. Such gaming techniques are available to certain social groups, i.e. the most digitally fluent ones, leading to wider disparities.[336] In parallel, revealing the workings of an AI system could expose other individuals' data, which would compromise their right to privacy, or national security

[332] Information Commissioner's Office, 'Big Data, Artificial Intelligence, Machine Learning and Data Protection'.
[333] Article 29 Data Protection Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679'.
[334] Andy Greenberg, 'An AI That Reads Privacy Policies So That You Don't Have To', *Wired*, 9 February 2018, https://www.wired.com/story/polisis-ai-reads-privacy-policies-so-you-dont-have-to/.
[335] Diakopoulos, 'Algorithmic Accountability'.
[336] Martin, 'Ethical Implications and Accountability of Algorithms'.

vulnerabilities, if the AI is used in the public sector.[337] Hence, occultation is on a certain level desirable.

Apart from 'gaming' and security risks, transparency disclosures related to proprietary AI signify the revelation of trade secrets on behalf of AI development firms. Undoubtedly, corporate entities see this as an unwelcome provision, going against their legitimate interests, reputation, and autonomy. As a result, the property interests of a company are pitted against individuals' interests for fair, lawful, and transparent processing of their data. As a matter of fact, the commercially sensitive algorithm of COMPAS is kept secret, which means that, in contrast to human decision-making, defendants have no ability to access its model and challenge the process by which it calculated their score.[338] Pertaining to such tensions, Recital 63 holds that data subjects' right to access their data should be satisfied to the extent that it does not seriously impair the rights or freedoms of others, such as trade secrets or intellectual property rights for software systems, albeit without leading to its complete refusal. However, WP29 contends that only under rare circumstances should such considerations override data subjects' rights.[339] Article 5 (d) of the Trade Secrets Directive also lifts its protective requirements to protect legitimate interests recognised by EU law, with data subjects' interests arguably belonging among these.[340] By the same token, Citron and Pasquale call for setting transparency expectations as the default status and permitting secrecy by way of exception, in lieu of the hitherto applied reverse model.[341] Ultimately, given that the rights and interests of both sides need to be considered in context, it is ambiguous how the appropriate balance will be stricken.

On the other side, Zerilli et al. argue that we should be careful not to demand more sophisticated transparency from AI than what is required from human decision-makers, lest we hold double standards.[342] By showcasing the obscurity frequently tolerated in the ratiocinations of human decisions, even at critical settings, they rebut the unrealistic expectations that some scholars foster for transparent AI.[343] In search of a middle ground, transparency disclosures could encompass not necessarily the innards of AI systems but the due process followed in

[337] Mittelstadt et al., 'The Ethics of Algorithms'.
[338] 'State v. Loomis'.
[339] Article 29 Data Protection Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679'.
[340] 'Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the Protection of Undisclosed Know-How and Business Information (Trade Secrets) against Their Unlawful Acquisition, Use and Disclosure', Pub. L. No. 32016L0943, OJ L 157 (2016), http://data.europa.eu/eli/dir/2016/943/oj/eng.
[341] Citron and Pasquale, 'The Scored Society: Due Process for Automated Predictions'.
[342] Zerilli et al., 'Transparency in Algorithmic and Human Decision-Making'.
[343] Zerilli et al.

their development and operation.[344] This transparency of due process or procedural regularity would reassure consumers that ethical and legal safeguards have been consistently observed in the processes followed by the AI system to generate its model and reach its conclusions, without jeopardising the company's commercial viability.[345]

Overall, although without a minimum degree of transparency individuals are deprived of the ability to monitor AI, transparency is not a panacea. Moreover, even if there was consensus on the content of transparency requirements, it is still obscure how AI developers could operationalise them in an AI system.

3.8 Auditing, certification, and enforcement

In its Briefing, the EPRS recommended the engagement of independent auditors, software watchdogs, or regulators, who would be responsible for investigating AI decisions.[346] Similarly, the WP29 endorses the adoption of systems that audit algorithms and review automated decision-making processes on the grounds of their accuracy and relevance.[347] The AI HLEG and Mittelstadt et al. concur in that requirements of explainability and transparency will be more effective if the relevant explanations or transparency disclosures are addressed not to lay data subjects but to third parties operating in the public interest.[348] External regulators and empirical researchers conducting audit or ethnographic studies are suitable to enforce algorithmic auditing.[349]

By accessing immutably recorded audit trails, third parties could verify at least key features of the performance of AI. Citron maintains that the minimum content of such algorithmic audit trails should consist of a full chronicle of conclusions reached in a particular case, the ways in which the facts considered in decision-making were assessed by human agents in addition to the identities of these human agents, and the rules followed by the AI system to form interim decisions.[350] Zerilli et al. discern two possible models of auditing: performance-based auditing, which would examine the functions of AI systems in view of their

[344] Martin, 'Ethical Implications and Accountability of Algorithms'.
[345] Citron and Pasquale, 'The Scored Society: Due Process for Automated Predictions'.
[346] Reillon, 'Understanding Artificial Intelligence'.
[347] Article 29 Data Protection Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679'.
[348] The High-Level Expert Group on Artificial Intelligence, 'Outcomes of the AI HLEG Workshop of 20 September 2018' (Brussels: Directorate-General for Communications Networks, Content and Technology, European Commission, 20 September 2018); Mittelstadt et al., 'The Ethics of Algorithms'.
[349] Mittelstadt et al., 'The Ethics of Algorithms'.
[350] Citron, 'Technological Due Process'.

outcomes; and accreditation-based auditing, which would assess the 'expertise' of the AI system, in the sense of possible credentials.[351] Thus, algorithmic auditing–much like audits in accounting and finance–is deemed as a good practice, used to ensure that AI does not reach discriminatory, biased, or otherwise askew conclusions and thereby offer public assurance, albeit without disclosing trade secrets.[352]

However, the efficiency of such auditing efforts would inevitably depend on the extent to which AI developers have designed their data-sets and algorithms as amenable to review. In the EU establishment, 'augmented public officials' such as judges, police, or European Commission staff using advanced digital tools are urged to lead by example in incorporating automated decision-making systems whose algorithms will be open to public auditing, testing, and review.[353] Auxiliary work-arounds to incentivise the development of audit-friendly algorithms would be for the EU to exclude providers of black-boxed AI systems from procurement or grant tax reliefs to those building compliant by design algorithms.[354]

To supplement auditing efforts, certification mechanisms, codes of practice, and ethical review boards could be introduced.[355] Indeed, Articles 40, 42, 43 and Recital 100 of the GDPR propose the voluntary implementation of codes of conduct, data protection certification schemes, and data protection seals or marks. Such credentials would indicate compliance with the GDPR and enhance transparency. Similarly, Martin suggests that, owing to the increased involvement of their products in crucial decision-making, developers should be under the oversight of authorities, which would grant accreditations to programmers conditional upon the completion of technical as well as ethical training.[356] Along the same lines, the EPSC suggests the extension of a Hippocratic Oath from traditional medical settings to the AI development field, as the latter gradually resembles medicine in terms of concerns of human well-being.[357] Given that for the time being the AI field is relatively freewheeling, guided by the 'move fast and break things' mentality of technological companies, for such measures to

[351] Zerilli et al., 'Transparency in Algorithmic and Human Decision-Making'.
[352] Information Commissioner's Office, 'Big Data, Artificial Intelligence, Machine Learning and Data Protection'.
[353] European Political Strategy Centre, 'The Age of Artificial Intelligence: Towards a European Strategy for Human-Centric Machines'.
[354] Ansgar Koene, 'Presentation of "a Governance Framework for Algorithmic Accountability and Transparency" at the European Parliament', UnBias, 6 November 2018, https://unbias.wp.horizon.ac.uk/2018/11/06/presentation-of-a-governance-framework-for-algorithmic-accountability-and-transparency-at-the-european-parliament/.
[355] Article 29 Data Protection Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679'.
[356] Martin, 'Ethical Implications and Accountability of Algorithms'.
[357] European Political Strategy Centre, 'The Age of Artificial Intelligence: Towards a European Strategy for Human-Centric Machines'.

be implemented its internal culture would need to undergo a thorough turnaround as well as a process of standardisation.

In the effort to monitor the execution of automated processing, national data protection authorities can be significantly helpful, since they are endowed with new enforcement powers following Chapters 6 and 8 of the GDPR. Data controllers should be particularly diligent in observing the relevant provisions. Non-compliance with the requirements of DPIAs, in the form of either omission or inadequate execution, entails fining up to 10 million Euros or 2% of the total worldwide annual turnover (Article 83 (4) a)). Critically, Article 83 (5) a) and b) ordains administrative fines up to 20 million Euros or 4% of the total worldwide annual turnover for non-compliance with the principles of data processing or the provisions of Article 22. Altogether, their investigatory and corrective powers in conjunction with their ability to impose steep fines reinforce their administrative status, overriding their previous perception as merely 'toothless data watchdogs'.[358]

3.9 Conclusion

All things considered, the AI, ML, and Big Data symbiosis, due to its distinct technical and ethical characteristics, is suboptimally subject to the GDPR provisions, which makes the relevance of the latter questionable. Without the necessary, clearly communicated, collective safeguards at place, data subjects have every reason to exhibit diminished trust and, consequently, diminished purchasing or usage interest in AI. In other words, the shrinkage of trustful consumer relations with AI will evenly wither market incentives for its development.[359] Contrariwise, the Directorate-General for Justice and Consumers noted that as more European citizens lose confidence in emerging technologies in the wake of recent data breaches (e.g. the Cambridge Analytica scandal), the high standards of data protection in the EU are credited with generating trust among consumers.[360] Contra the European Parliament's unrealistic claim that rules on intelligent machines must not affect their development, such technologies not only

[358] Bryan Casey, Ashkon Farhangi, and Roland Vogl, 'Rethinking Explainable Machines: The GDPR's "Right to Explanation" Debate and the Rise of Algorithmic Audits in Enterprise', *Berkeley Technology Law Journal* Forthcoming (19 February 2018), https://ssrn.com/abstract=3143325.
[359] Tal Zarsky, 'The Privacy–Innovation Conundrum', *Lewis & Clark Law Review* 19, no. 1 (20 April 2015): 115–68.
[360] Věra Jourová, 'The EU Data Protection Reform and Big Data', Factsheet (Directorate-General for Justice and Consumers, European Commission, January 2016), http://ec.europa.eu/newsroom/just/document.cfm?doc_id=41523.

influence regulation, but, in reverse, regulation inevitably influences their future surge and adoption.[361]

On a more positive note, a holistic leverage of the principles and rights enshrined in the GDPR along with the enforcement of DPIAs and algorithmic auditing is in a right–yet timid–course to empower data subjects with remedial means against high-tech discrimination. Moving further, the EU should examine whether this approach could be enhanced by binding measures concretely targeted to AI.

[361] European Parliament et al., 'Civil Law Rules on Robotics'.

Conclusion

The EU response and its evaluation

At the outset of this dissertation, the topicality of biased AI came to the fore through stories of AI-enabled discrimination which struck a public nerve in the US and the EU. This dissertation adopted an interdisciplinary orientation to answer which are the main ethical and legal challenges that the EU faces with respect to biased and discriminatory Narrow AI, especially of the ML type fuelled by (Big) Data.

Delving into the technical background of AI and its interlocking ML and Big Data technologies was imperative at the beginning of this analysis, as ethicists and legal practitioners alike need to have a solid grasp of the technological manifestations they are evaluating. At the same time, the adoption of AI latches onto the adoption of well-informed, rational approaches thereto on behalf of society. This is why a working definition of AI was offered, based on the core conditions found in EU communications and with the hope that EU bodies will also conclude to a stable definition balancing between accuracy and flexibility. A historical perspective likewise contributed to situating readers in the AI trajectory and forming reasonable expectations for the future. As the possibility and, a fortiori, the threats of General AI are far-off given the current state-of-the-art, the remainder of this dissertation focused on the most recent and widespread type of AI, namely Narrow AI, of the ML subset and grounded in Big Data. Among the technical features of this type of AI, its self-learning abilities and black-box operation emerged as causes of vexing challenges in the ethical and legal space.

Specifically, the deployment of AI does not exonerate the network of material and human agents wherein it acts from the attribution of moral responsibility. Even when in working order, AI systems are morally biased, as they favour particular individuals and groups and unfairly discriminate against others. Against empiricist claims of objectivity, there is a strong case to be made that AI systems are sociotechnical artefacts riddled with subjective evaluations. In selecting the outcomes of an AI system and the categories into which they will classify people, AI developers rely on their biased understanding and arbitrary perceptions of the problem at stake. Moving on to the training phase, the data from which AI systems learn are affected by the positionality of those handling them or misrepresent different social groups. Simultaneously, by attributing normative dimensions to descriptive representations of reality, AI systems succumb to the naturalistic fallacy and preserve historical injustices. Except for the dubious quality of data and their manipulation, AI often discriminates against protected social

groups based on seemingly neutral proxies and unveils misleading correlations. Once deployed, the effects of AI systems are unfair on egalitarian grounds. In allocating valuable resources, on the one hand, they include considerations about people's lives which are attributed to luck and, on the other hand, they exclude considerations of the historical and social context wherein their data were generated. Such allocative harms are often preceded by representative harms.

Trying to address instances of biased AI, individuals face barriers because of their limited technical understanding compared to that of AI developers, which results in a perceptible power imbalance. Yet, the sophisticated technical understanding of AI developers only goes so far. The inherent inexplicability and constant modification of the internal mechanisms of AI impedes even its developers from fully comprehending how it reaches its decisions, whereas it can be used as an excuse to mask deliberate discriminatory or generally illegitimate practices. In sum, the ethical neutrality of these data-rich tools is refuted because of epistemological ethical concerns, in the sense that they produce unjustified beliefs through unreliable processes, and normative ethical concerns, in the sense that they oppose norms of luck egalitarianism and deontic egalitarianism.

The exposition of ethical concerns provides a segue into nascent tensions on the legal frontier, wherein biased AI decisions can have significant legal consequences, such as causing one's incarceration or incriminating them for illegal entry to a country. The EU response to biased AI surfaces in two ways. Firstly, the applicability of GDPR to data-driven AI is flagged up, making it necessary for the European Data Protection Board and national data protection authorities to become familiar with such technologies. Secondly, the advancement of AI has led part of EU policy-makers to express a need for legislative adjustments based upon the EU's fundamental values and principles.

This need for legal reforms becomes clear in examining the tensions between AI and the GDPR. Specifically, AI is irreconcilable with the principle of purpose limitation, as it finds patterns and correlations among data on its own, with the purpose of such analysis becoming clear only after its occurrence. It is likewise irreconcilable with the principle of data minimisation, given that its function is supported by utilising all available data, beyond the merely necessary and relevant ones. Its reliance on Big Data, proved to be messy and misrepresentative, clashes with the principle of accuracy and necessitates their replacement with better quality data, sourced from diversified teams or the public sector. Even if AI uses accurate data, though, it tends to store them for unrestricted periods, running counter to the principle of storage limitation. On the other side, DPIAs, under the principle of accountability,

do not guarantee sufficiently substantive protection against unfair discrimination. Bringing the data subjects' perspective forward, it is difficult to meet the threshold of the right not to be subject to automated decision-making, a right which on the whole creates more uncertainty than it resolves. Their inability to grant informed consent to unpredictable processing alongside the prohibition of solely automated decision-making and special category data processing limit the legal bases of processing dictated by the principle of lawfulness, whereas the principle of fairness can only come to bear if it incorporates collective and representational harms. Compliance with both principles calls for the principle of transparency, which is useful in assessing the due process followed by AI but hard to align with risks of manipulation, security threats, and intellectual property disclosures. Hopefully, auditing and certification mechanisms seem promising, yet dependent on the extent to which AI developers design their systems as amenable to review and are eager to transform the culture of their industry. Similarly, the enhanced powers of national data protection authorities are expected to have deterrent effects to negligent data processing.

By bringing together insights drawn from the ethical and legal aspects of AI bias, their cumulative effect nurtures a deeper understanding and distils the following concluding remarks on the road to bridging the 'pacing problem of law'.

Firstly, the ethical and legal argumentation coincide in that we should acknowledge the vast gamut of data-driven AI applications and pay special attention to those taking final decisions, which deeply affect individuals' activities and interests, for instance in criminal justice or policing, with implications diffusing to their family and social circle. Just because many decisions can be automated by AI does not necessarily mean that they should be so or that they should all be treated the same; instead, there should be clear red lines when developing and deploying AI. In contexts crossing such red lines, AI systems need not be abolished but at least limited to an advisory, interim step supporting human decision-makers who will be responsible for the ultimate conclusion. Simply put, in some procedures the stakes are so high that there should always be a 'human-in-the-loop'.[362] More broadly, the greater the impact of the decision to be made, the higher epistemic, normative, and legal justifications should be required.

Secondly, anticipatory approaches such as value-sensitive design in the realm of ethics and DPIAs, built-in features for auditing, or similar ex ante measures in the realm of law

[362] Human-in-the-loop means the active, often continuous, engagement of a human in control decisions: William D. Nothwang et al., 'The Human Should Be Part of the Control Loop?', in *2016 Resilience Week (RWS)* (2016 Resilience Week (RWS), Chicago, USA: IEEE, 2016), 214–20, https://doi.org/10.1109/RWEEK.2016.7573336.

encounter less friction with AI. As they acknowledge the values and risks inherent in AI development, they forestall discrimination before data subjects are harmed and undeservedly burdened to prove so. Given that such approaches are more likely to be effective against bias-busting, they could serve as indicators of quality AI development and become indispensable requirements in forthcoming research proposals.[363]

Thirdly, the principles-based approach of Article 5 rightfully identifies critical areas in data-driven technologies and offers flexibility when the details of problematic cases are not yet brought to light. Nonetheless, when it comes to practical compliance with these principles or cognate provisions and their operationalisation in AI systems, the technical peculiarities of data-driven ML models spark unresolved tensions vis-à-vis the Regulation. The interpretative ambiguity and dearth of clarity in the GDPR make matters worse. What could be a powerful tool in the hands of data subjects, namely a right to explanation, is still under controversy because of this exact ambiguity, whereas the prohibition of automated decision-making presupposes the satisfaction of so many vague conditions that its application might end up feeble. At the same time, a paradox surfaces: whereas the GDPR aims at empowering data subjects by restricting the use of their data, AI needs more, if not all, data to form accurate and representative training data-sets and prevent biases from seeping into them. All these factors lend weight to the existence of AI-related legal gaps in the GDPR, which should be indicatively addressed through clear interpretations or supplementary sui generis provisions. To avoid a premature overregulation of still under-examined issues, a course of action aligned with the EU's Better Regulation Agenda would be the implementation of regulatory sandboxes, meaning testing environments to facilitate experimentation with AI under a regulator's supervision and identify pain points in the current legislation.[364]

Lastly, having taken stock of the ethical and legal challenges of biased AI, it becomes apparent that a well-thought-out EU policy response would not hinder AI development, as critics might support, but would rather reinstate the already eroded public trust towards AI, ML, and Big Data. Such policies could reinforce the integration of AI into private and public life and become EU's competitive advantage in the global landscape. As a result, boosting AI

[363] Directorate-General for Communications Networks, Content and Technology, 'Coordinated Plan on Artificial Intelligence', Annex to the Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions (Brussels: European Commission, 7 December 2018), https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56017.

[364] Directorate-General for Communications Networks, Content and Technology; Francesca Jenner, 'Better Regulation: Why and How', European Commission - European Commission, 30 May 2016, https://ec.europa.eu/info/law/law-making-process/planning-and-proposing-law/better-regulation-why-and-how_en.

development and guaranteeing protection against unwanted bias do not stand in an either-or relationship, but the latter facilitates the former. As the issue of trust recurred at several junctures of this dissertation, the telos of such policy endeavours should be a Trustworthy AI. The ideal of Trustworthy AI means that when AI systems are not epistemically accessible in their entirety, the concerned parties should be confident that the result achieved will not be necessarily favourable but at least compliant with ex ante prescriptive standards and ex post abilities for redress. At their core, trustful relationships either with machines or humans entail risk-taking: it is a leap of faith, albeit with choices and safeguards in place.

Trust in AI, based on respect to fundamental rights and ethical rules, is one of the cornerstones of the Commission's *Coordinated Plan on Artificial Intelligence*.[365] Along similar lines, the AI HLEG structured its initial discussions under the overarching theme of Trustworthy AI, which results from combining laws, standards, and certifications with human-oriented, ethical considerations on an ongoing basis.[366] Although the AI HLEG did not define Trustworthy AI, ensuring an ethical intent aligned with core values and principles during AI development and implementing it with both technical and non-technical tools were identified as its foundational pillars.[367] These pillars saliently epitomise key aspects of the herein exposition and implicitly show that trustworthy algorithmic components are not enough for Trustworthy AI; the entire sociotechnical system wherein AI is deployed must be so.

Since an unrealistic desire to fully control emerging technologies would paralyse even daily actions, placing trust to keep data controllers accountable is also endorsed by O'Neill as an effective mechanism when meaningful consent, transparency, or explainability is not an option.[368] Instead of aiming at a complete eradication of AI-enabled risks, we need feasible ways to minimise them and enable data processing by private and public entities in order to reap the maximum benefits of AI. Thus, the ideal of Trustworthy AI could be an ethical and legal compass helping policy-makers navigate between the Scylla of unconditionally accepting efficient but inexplicable, potentially biased AI systems and the Charybdis of guaranteeing maximalist but innovation-strangling protection of individual and collective rights.

[365] Directorate-General for Communications Networks, Content and Technology, 'Coordinated Plan on Artificial Intelligence', 7 December 2018.
[366] The High-Level Expert Group on Artificial Intelligence, 'Outcomes of the AI HLEG Workshop of 20 September 2018'.
[367] The High-Level Expert Group on Artificial Intelligence.
[368] Tae Wan Kim and Bryan Routledge, 'Algorithmic Transparency, A Right To Explanation and Trust', June 2017, 31.

Recent initiatives

Keeping in mind the oft-quoted demand for interdisciplinarity to approximate the ideal of Trustworthy AI, the policy framework needs to dovetail with insights from the AI ecosystem.

Fairness, Accountability, and Transparency in Machine Learning (FAT-ML) is an international workshop convening prominent researchers, practitioners, and policy-makers to explore solutions to the ethical and legal conundrums of ML discrimination.[369] The *Principles for Accountable Algorithms and a Social Impact Statement for Algorithm* published by FAT-ML as well as *The Toronto Declaration* by a coalition of non-profits and technologists advance noteworthy recommendations to counteract discriminatory ML.[370] In a similar spirit, Joy Buolamwini, a black MIT researcher whose face was consistently unrecognised by AI, launched the Algorithmic Justice League (AJL).[371] The AJL exposes algorithmic bias and establishes a community where people share concerns and experiences of coded discrimination.[372] At an independent level, an ascendant number of researchers explore value sensitive design methods to integrate fairness and discrimination awareness into data-driven AI systems.[373]In parallel with investigative journalism, whose quick reflexes in detecting biased AI were showcased in the Introduction, and the newly spawned fields of algorithmic accountability reporting and information technology whistleblowing, such initiatives to 'debias' AI should be encouraged and borne in mind by policy-makers.[374]

Future directions

As the AI HLEG is expected to release its AI ethics guidelines and policy recommendations in March 2019 with input from the European AI Alliance, their insights will play a crucial role to

[369] 'Fairness, Accountability, and Transparency in Machine Learning', accessed 11 November 2018, http://www.fatml.org/.
[370] 'Principles for Accountable Algorithms and a Social Impact Statement for Algorithms: FAT ML', accessed 12 November 2018, http://www.fatml.org/resources/principles-for-accountable-algorithms; 'The Toronto Declaration: Protecting the Right to Equality and Non-Discrimination in Machine Learning Systems', 16 May 2018, https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf.
[371] 'AJL -Algorithmic Justice League', AJL, accessed 11 November 2018, https://www.ajlunited.org/.
[372] 'Project Overview Algorithmic Justice League', MIT Media Lab, accessed 11 November 2018, https://www.media.mit.edu/projects/algorithmic-justice-league/overview/.
[373] Cynthia Dwork et al., 'Fairness Through Awareness', *ArXiv:1104.3913 [Cs]*, 19 April 2011, http://arxiv.org/abs/1104.3913; Dino Pedreshi, Salvatore Ruggieri, and Franco Turini, 'Discrimination-Aware Data Mining', in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '08, Las Vegas, USA: ACM Press, 2008), 560, https://doi.org/10.1145/1401890.1401959. Cf. Andrew D. Selbst et al., 'Fairness and Abstraction in Sociotechnical Systems', in *ACM Conference on Fairness, Accountability, and Transparency*, vol. 1, 1 (FAT*, Rochester, USA, 2018), https://papers.ssrn.com/abstract=3265913.
[374] Diakopoulos, 'Algorithmic Accountability'.

resolving the challenges of biased AI in the EU. The Algorithmic Awareness Building project recently submitted its first draft report for peer review, so the evaluation of its results is deferred for future commentary.[375] At Member State level, a considerable number of them have published national AI strategies, others are in the process of doing so, and the rest are encouraged by the Commission to follow suit by mid-2019.[376] It would be interesting to explore comparatively whether these national strategies will include provisions on AI discrimination and, if yes, what their relation will be with the broader EU policy response, especially under the Commission's new term and the European Parliament's new composition in 2019. In the other direction, it is worth examining how EU-wide measures could be enhanced by policy responses at an international level, for instance under the G7 Summit. Overall, the design and implementation of AI-targeted legislation will remain at the core of discussion. However, as AI, ML, and Big Data steadily penetrate legal science, it would be thought-provoking to study how they could be employed to resolve their self-inflicted harms.[377]

On a final note, by acknowledging the limits of the convergent AI, ML, and Big Data technologies, this dissertation does not wish to paint just a grim picture of the problem and countenance their dismissal. Rather, it aspires to have enabled a better grounded and ultimately more valuable set of conversations about the ethically and legally informed development of such game-changing technologies. Questions about biases and discrimination are in all cases hard to tackle and far from being settled. Although the scalability and extolled objectivity of AI definitely exacerbate biases and discrimination, at the end of the day it is human agents who have embedded these biases and, thus, ought to remain vigilant and accountable with regard to their treatment. Hence, this dissertation wishes to serve as a stepping stone for readers and scholars to further reflect not only on the underlying reasons of biased Artificial Intelligence, but most importantly on the reasons why human intelligence and society have been ingrained with such biases in the first place.

[375] 'State of the Art Report', Algo Aware, accessed 7 December 2018, https://www.algoaware.eu/state-of-the-art-report/.
[376] Directorate-General for Communications Networks, Content and Technology, 'Coordinated Plan on Artificial Intelligence', 7 December 2018.
[377] See: Nikolaos Aletras et al., 'Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective', PeerJ Computer Science 2 (24 October 2016): e93, https://doi.org/10.7717/peerj-cs.93; Marios Koniaris, Ioannis Anagnostopoulos, and Yannis Vassiliou, 'Network Analysis in the Legal Domain: A Complex Model for European Union Legal Sources', Journal of Complex Networks 6, no. 2 (1 April 2018): 243–68, https://doi.org/10.1093/comnet/cnx029.

Bibliography

Primary sources

Legislation

Charter of Fundamental Rights of the European Union, Pub. L. No. 2007/C 303/01, OJ C 303 (2007). https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:12007P/TXT.

Consolidated versions of the Treaty on European Union and the Treaty on the Functioning of the European Union, Pub. L. No. 2008/C 115/01, OJ C 115 (2008). https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1544456032916&uri=CELEX:C2008/115/01.

Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin, Pub. L. No. 32000L0043, OJ L 180 (2000). http://data.europa.eu/eli/dir/2000/43/oj/eng.

Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation, Pub. L. No. 32000L0078, OJ L 303 (2000). http://data.europa.eu/eli/dir/2000/78/oj/eng.

Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services, Pub. L. No. 32004L0113, OJ L 373 (2004). http://data.europa.eu/eli/dir/2004/113/oj/eng.

Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, Pub. L. No. 31995L0046, OJ L 281 (1995). http://data.europa.eu/eli/dir/1995/46/oj/eng.

Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast), Pub. L. No. 32006L0054, OJ L 204 (2006). http://data.europa.eu/eli/dir/2006/54/oj/eng.

Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA, Pub. L. No. 32016L0680, OJ L 119 (2016). http://data.europa.eu/eli/dir/2016/680/oj/eng.

Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure, Pub. L. No. 32016L0943, OJ L 157 (2016). http://data.europa.eu/eli/dir/2016/943/oj/eng.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data

Protection Regulation), Pub. L. No. 32016R0679, OJ L 119 (2016). http://data.europa.eu/eli/reg/2016/679/oj/eng.

Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union, Pub. L. No. 32018R1807, OJ L 303 (2018). http://data.europa.eu/eli/reg/2018/1807/oj/eng.

Reports, guidelines, opinions & other EU communications

'AI Forum 2018 in Finland'. Digital Single Market. Accessed 24 November 2018. https://ec.europa.eu/digital-single-market/en/news/ai-forum-2018-finland.

'AI Watch'. Knowledge for policy, 28 November 2018. https://ec.europa.eu/knowledge4policy/ai-watch_en.

Article 29 Data Protection Working Party. 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679', 3 October 2017. http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053.

———. 'Guidelines on Data Protection Impact Assessment (DPIA) and Determining Whether Processing Is "Likely to Result in a High Risk" for the Purposes of Regulation 2016/679', 4 October 2017. http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611236.

———. 'Opinion 03/2013 on Purpose Limitation', 2 April 2013. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf.

———. 'Statement on the Role of a Risk-Based Approach in Data Protection Legal Frameworks', 30 May 2014. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp218_en.pdf.

Bentley, Peter, Miles Brundage, Olle Häggström, and Thomas Metzinger. *Should We Fear Artificial Intelligence?: In-Depth Analysis*. Brussels: Scientific Foresight Unit, European Parliament, 2018. http://www.europarl.europa.eu/RegData/etudes/IDAN/2018/614547/EPRS_IDA(2018)614547_EN.pdf.

'Commissioner Gabriel Hosted First High-Level Forum of Member States on Digitalisation of Industry and Artificial Intelligence'. Digital Single Market. Accessed 24 November 2018. https://ec.europa.eu/digital-single-market/en/blogposts/commissioner-gabriel-hosted-first-high-level-forum-member-states-digitalisation-industry.

Delvaux, Mady. 'Report with Recommendations to the Commission on Civil Law Rules on Robotics'. Strasbourg: Committee on Legal Affairs, European Parliament, 27 January 2017. http://www.europarl.europa.eu/sides/getDoc.do?type=REPORT&reference=A8-2017-0005&language=EN.

Directorate-General for Communications Networks, Content and Technology. 'Artificial Intelligence for Europe'. Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the regions. Brussels: European Commission, 25

April 2018. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237&rid=2.

———. 'Coordinated Plan on Artificial Intelligence'. Annex to the Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. Brussels: European Commission, 7 December 2018. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56017.

———. 'Coordinated Plan on Artificial Intelligence'. Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. Brussels: European Commission, 7 December 2018. https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1544192414694&uri=COM:2018:795:FIN.

Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs. 'On the Road to Automated Mobility: An EU Strategy for Mobility of the Future'. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee, the Committee of the Regions. Brussels: European Commission, 17 May 2018. https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1541878088823&uri=CELEX:52018DC0283.

European Commission. 'Artificial Intelligence: Commission Kicks off Work on Marrying Cutting-Edge Technology and Ethical Standards', 3 September 2018. http://europa.eu/rapid/press-release_IP-18-1381_en.htm.

European Data Protection Board. 'Endorsement 1/2018'. Accessed 11 November 2018. https://edpb.europa.eu/sites/edpb/files/files/news/endorsement_of_wp29_documents.pdf.

European Group on Ethics in Science and New Technologies. 'Artificial Intelligence, Robotics and "Autonomous" Systems'. Statement. Luxembourg: Directorate-General for Research and Innovation, European Commission, 30 April 2018. https://publications.europa.eu/en/publication-detail/-/publication/dfebe62e-4ce9-11e8-be1d-01aa75ed71a1/language-en.

European Parliament, Committee on Civil Liberties, Justice and Home Affairs, Committee on Employment and Social Affairs, Committee on Industry, Research and Energy, Committee on Legal Affairs, Committee on Transport and Tourism, Committee on the Environment, Public Health and Food Safety, and Committee on the Internal Market and Consumer Protection. 'European Parliament Resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))'. *Official Journal of the European Union*, C, 61, no. 252 (18 July 2018): 239–257.

European Political Strategy Centre. 'The Age of Artificial Intelligence: Towards a European Strategy for Human-Centric Machines'. EPSC Strategic Notes. European Commission, 27 March 2018. https://ec.europa.eu/epsc/sites/epsc/files/epsc_strategicnote_ai.pdf.

European Union Agency for Fundamental Rights. '#BigData: Discrimination in Data-Supported Decision Making'. Vienna, 29 May 2018. http://fra.europa.eu/en/publication/2018/big-data-discrimination.

FET Advisory Group. 'The Need to Integrate the Social Sciences and Humanities with Science and Engineering in Horizon 2020 and Beyond'. European Commission,

December 2016. https://ec.europa.eu/digital-single-market/en/news/report-need-integrate-social-sciences-and-humanities-science-and-engineering-horizon-2020.

Guerini, Giuseppe. 'Artificial Intelligence for Europe (Communication)'. Opinion. Brussels: European Economic and Social Committee, 19 September 2018. https://www.eesc.europa.eu/en/our-work/opinions-information-reports/opinions/artificial-intelligence-europe-communication.

'Human Brain Project Flagship'. Digital Single Market. Accessed 4 November 2018. https://ec.europa.eu/digital-single-market/en/human-brain-project.

Jenner, Francesca. 'Better Regulation: Why and How'. European Commission - European Commission, 30 May 2016. https://ec.europa.eu/info/law/law-making-process/planning-and-proposing-law/better-regulation-why-and-how_en.

Jourová, Věra. 'The EU Data Protection Reform and Big Data'. Factsheet. Directorate-General for Justice and Consumers, European Commission, January 2016. http://ec.europa.eu/newsroom/just/document.cfm?doc_id=41523.

Madiega, Tambiama. 'Copyright in the Digital Single Market'. Briefing. European Parliamentary Research Service, European Parliament, July 2018. http://www.europarl.europa.eu/RegData/etudes/BRIE/2016/593564/EPRS_BRI(2016)593564_EN.pdf.

Muller, Catelijne. 'Artificial Intelligence – The Consequences of Artificial Intelligence on the (Digital) Single Market, Production, Consumption, Employment and Society'. Opinion. Brussels: European Economic and Social Committee, 31 May 2017. https://www.eesc.europa.eu/en/our-work/opinions-information-reports/opinions/artificial-intelligence.

Negreiro, Mar. 'Re-Use of Public Sector Information'. Briefing. European Parliamentary Research Service, European Parliament, November 2018. http://www.europarl.europa.eu/RegData/etudes/BRIE/2018/628312/EPRS_BRI(2018)628312_EN.pdf.

Probst, Laurent, Bertrand Pedersen, Virginie Lefebvre, and Lauriane Dakkak-Arnoux. 'USA-China-EU Plans for AI: Where Do We Stand?' Digital Transformation Monitor, European Commission, January 2018. https://ec.europa.eu/growth/tools-databases/dem/monitor/sites/default/files/DTM_AI%20USA-China-EU%20plans%20for%20AI%20v5.pdf.

'Processing Power beyond Moore's Law | News'. CORDIS | European Commission. Accessed 4 November 2018. https://cordis.europa.eu/news/rcn/129281_en.html.

Reillon, Vincent. 'Understanding Artificial Intelligence'. Briefing. European Parliamentary Research Service, European Parliament, January 2018. http://www.iberglobal.com/files/2018/Understanding_AI.pdf.

Secretariat-General. 'A Connected Digital Single Market for All'. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on the Mid-Term Review on the implementation of the Digital Single Market Strategy. Brussels: European Commission, 5 October 2017. https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1511969120337&uri=CELEX:52017DC0228&print=true.

'State of the Art Report'. Algo Aware. Accessed 7 December 2018. https://www.algoaware.eu/state-of-the-art-report/.

Stix, Charlotte. ‘The European AI Landscape’. Brussels: Directorate-General for Communications Networks, Content and Technology, European Commission, 18 April 2018. https://ec.europa.eu/digital-single-market/en/news/european-artificial-intelligence-landscape.

The High-Level Expert Group on Artificial Intelligence. ‘Outcomes of the AI HLEG Workshop of 20 September 2018’. Brussels: Directorate-General for Communications Networks, Content and Technology, European Commission, 20 September 2018.

Waterbley, Séverine, Ivan Dimov, Ondřej Malý, Søren Gaard, Georges Friden, József Pálinkás, Evarist Bartolo, et al. ‘Cooperation on Artificial Intelligence’. Declaration, 4 October 2018. https://ec.europa.eu/digital-single-market/en/news/eu-member-states-sign-cooperate-artificial-intelligence.

Case law

Judgment of the Court (Grand Chamber). Case C-236/09, Association belge des Consommateurs Test-Achats ASBL and Others v. Conseil des ministres (European Court of Justice 1 March 2011).

‘State v. Loomis’. Harvard Law Review. Accessed 12 November 2018. https://harvardlawreview.org/2017/03/state-v-loomis/.

Secondary sources

Books

Asimov, Isaac. *I, Robot*. Accessed 11 October 2018. https://www.ttu.ee/public/m/mart-murdvee/Techno-Psy/Isaac_Asimov_-_I_Robot.pdf.

‘Homer, Iliad, Book 18’. Accessed 11 October 2018. http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.01.0134%3Abook%3D18.

Latour, Bruno, Steve Woolgar, and Jonas Salk. *Laboratory Life: The Construction of Scientific Facts*. Princeton, United States: Princeton University Press, 1986. http://ebookcentral.proquest.com/lib/kcl/detail.action?docID=1144731.

Marchant, Gary E., Braden R. Allenby, and Joseph R. Herkert. *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem*. Dordrecht: Springer Science+Business Media B.V., 2011. http://0-dx.doi.org.fama.us.es/10.1007/978-94-007-1356-7.

Mayer-Schönberger, Viktor, and Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston, MA: Houghton Mifflin Harcourt, 2013.

Pasquale, Frank. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, United States: Harvard University Press, 2015. http://ebookcentral.proquest.com/lib/kcl/detail.action?docID=3301535.

Russell, Stuart J., Peter Norvig, and Ernest Davis. *Artificial Intelligence: A Modern Approach*. 3rd ed. Prentice Hall Series in Artificial Intelligence. Upper Saddle River: Prentice Hall, 2010.

Sidgwick, Henry. *Outlines of the History of Ethics for English Readers*. London: Macmillan, 1886. http://archive.org/details/outlinesofhistor00sidguoft.

'The Argonautica, by Apollonius Rhodius'. Accessed 11 October 2018. http://www.gutenberg.org/files/830/830-h/830-h.htm.

Articles & papers

Aletras, Nikolaos, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 'Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective'. *PeerJ Computer Science* 2 (24 October 2016): e93. https://doi.org/10.7717/peerj-cs.93.

Ananny, Mike. 'Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness'. *Science, Technology, & Human Values* 41, no. 1 (January 2016): 93–117. https://doi.org/10.1177/0162243915606523.

Andersen, Hanne, and Brian Hepburn. 'Scientific Method'. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2016. Metaphysics Research Lab, Stanford University, 2016. https://plato.stanford.edu/archives/sum2016/entries/scientific-method/.

Anderson, Chris. 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete'. *Wired*, 23 June 2008. https://www.wired.com/2008/06/pb-theory/.

Angwin, Julia. 'Sample COMPAS Risk Assessment'. Accessed 19 November 2018. https://www.propublica.org/documents/item/2702103-Sample-Risk-Assessment-COMPAS-CORE.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 'Machine Bias'. ProPublica, 23 May 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

'Artificial Intelligence Is Awakening the Chip Industry's Animal Spirits'. *The Economist*, 7 June 2018. https://www.economist.com/business/2018/06/07/artificial-intelligence-is-awakening-the-chip-industrys-animal-spirits.

B. Solum, Lawrence. 'Legal Personhood for Artificial Intelligences'. *North Carolina Law Review* 70, no. 4 (1 April 1992): 1231–87.

Bano, Muneera. 'Artificial Intelligence Is Demonstrating Gender Bias – and It's Our Fault', 25 July 2018. https://www.kcl.ac.uk/news/news-article.aspx?id=c97f7c12-ae02-4394-8f84-31ba4d56ddf7.

Barocas, Solon, and Andrew D. Selbst. 'Big Data's Disparate Impact'. *California Law Review* 104 (2016): 671–732. https://doi.org/10.15779/z38bg31.

Binns, Reuben. 'Data Protection Impact Assessments: A Meta-Regulatory Approach'. *International Data Privacy Law* 7, no. 1 (February 2017): 22–35. https://doi.org/10.1093/idpl/ipw027.

———. 'Fairness in Machine Learning: Lessons from Political Philosophy'. In *Proceedings of Machine Learning Research*, 81:1–11, 2017. http://arxiv.org/abs/1712.03586.

Boffey, Daniel. 'EU Border "lie Detector" System Criticised as Pseudoscience'. *The Guardian*, 2 November 2018, sec. World news.

https://www.theguardian.com/world/2018/nov/02/eu-border-lie-detection-system-criticised-as-pseudoscience.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 'Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings'. *ArXiv:1607.06520 [Cs, Stat]*, 21 July 2016. http://arxiv.org/abs/1607.06520.

boyd, danah, and Kate Crawford. 'Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon'. *Information, Communication & Society* 15, no. 5 (June 2012): 662–79. https://doi.org/10.1080/1369118X.2012.678878.

Buolamwini, Joy, and Timnit Gebru. 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification'. In *Proceedings of Machine Learning Research*, 81:77–91, 2018.

Burrell, Jenna. 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms'. *Big Data & Society* 3, no. 1 (5 January 2016): 1–12. https://doi.org/10.1177/2053951715622512.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 'Semantics Derived Automatically from Language Corpora Contain Human-like Biases'. *Science* 356, no. 6334 (14 April 2017): 183–86. https://doi.org/10.1126/science.aal4230.

Calude, Cristian S., and Giuseppe Longo. 'The Deluge of Spurious Correlations in Big Data'. *Foundations of Science* 22, no. 3 (September 2017): 595–612. https://doi.org/10.1007/s10699-016-9489-4.

Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 'Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission'. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–30. Sydney, Australia: ACM Press, 2015. https://doi.org/10.1145/2783258.2788613.

Casey, Bryan, Ashkon Farhangi, and Roland Vogl. 'Rethinking Explainable Machines: The GDPR's "Right to Explanation" Debate and the Rise of Algorithmic Audits in Enterprise'. *Berkeley Technology Law Journal* Forthcoming (19 February 2018). https://ssrn.com/abstract=3143325.

Cave, Stephen, and Kanta Dihal. 'Ancient Dreams of Intelligent Machines: 3,000 Years of Robots'. *Nature* 559 (25 July 2018): 473. https://doi.org/10.1038/d41586-018-05773-y.

Chander, Anupam. 'The Racist Algorithm?' *Michigan Law Review* 115, no. 6 (2017): 1023–45.

Chavalarias, David, and John P.A. Ioannidis. 'Science Mapping Analysis Characterizes 235 Biases in Biomedical Research'. *Journal of Clinical Epidemiology* 63, no. 11 (November 2010): 1205–15. https://doi.org/10.1016/j.jclinepi.2009.12.011.

Citron, Danielle Keats. 'Technological Due Process'. *Washington University Law Review* 85, no. 6 (2008): 1249–1313.

Citron, Danielle Keats, and Frank Pasquale. 'The Scored Society: Due Process for Automated Predictions'. *Washington Law Review*, University of Maryland Legal Studies Research Paper No. 2014-8, 89 (2014): 1–33.

Crawford, Kate. ‘The Hidden Biases in Big Data’. *Harvard Business Review*, 1 April 2013. https://hbr.org/2013/04/the-hidden-biases-in-big-data.

Danaher, John. ‘The Threat of Algocracy: Reality, Resistance and Accommodation’. *Philosophy & Technology* 29, no. 3 (September 2016): 245–68. https://doi.org/10.1007/s13347-015-0211-1.

Dastin, Jeffrey. ‘Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women’. *Reuters*, 10 October 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Delacroix, Sylvie. ‘Pervasive Data Profiling, Moral Equality and Civic Responsibility’. *SSRN Electronic Journal*, 2017. https://doi.org/10.2139/ssrn.3022188.

Diakopoulos, Nicholas. ‘Algorithmic Accountability: Journalistic Investigation of Computational Power Structures’. *Digital Journalism* 3, no. 3 (4 May 2015): 398–415. https://doi.org/10.1080/21670811.2014.976411.

Dietterich, Thomas G., and Eun Bae Kong. ‘Machine Learning Bias, Statistical Bias, and Statistical Variance of Decision Tree Algorithms’. Oregon State University, 1995. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.38.2702&rep=rep1&type=pdf.

Dobbe, Roel, Sarah Dean, Thomas Gilbert, and Nitin Kohli. ‘A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics’. In *2018 Workshop on Fairness, Accountability, and Transparency in Machine Learning*. Stockholm, Sweden, 2018. http://arxiv.org/abs/1807.00553.

Dressel, Julia, and Hany Farid. ‘The Accuracy, Fairness, and Limits of Predicting Recidivism’. *Science Advances* 4, no. 1 (January 2018): eaao5580. https://doi.org/10.1126/sciadv.aao5580.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. ‘Fairness Through Awareness’. *ArXiv:1104.3913 [Cs]*, 19 April 2011. http://arxiv.org/abs/1104.3913.

Elish, M. C., and danah boyd. ‘Situating Methods in the Magic of Big Data and AI’. *Communication Monographs* 85, no. 1 (2 January 2018): 57–80. https://doi.org/10.1080/03637751.2017.1375130.

Finley, Taryn. ‘Google Apologizes for Tagging Photos of Black People as “Gorillas”’. HuffPost UK, 2 July 2015. http://www.huffingtonpost.com/2015/07/02/google-black-people-goril_n_7717008.html.

Floridi, Luciano, Mariarosaria Taddeo, and Matteo Turilli. ‘Turing’s Imitation Game: Still an Impossible Challenge for All Machines and Some Judges—An Evaluation of the 2008 Loebner Contest’. *Minds and Machines* 19, no. 1 (February 2009): 145–50. https://doi.org/10.1007/s11023-008-9130-6.

Ford, Heather. ‘Big Data and Small: Collaborations between Ethnographers and Data Scientists’. *Big Data & Society* 1, no. 2 (10 July 2014): 1–3. https://doi.org/10.1177/2053951714544337.

Forsythe, Diana E. ‘Engineering Knowledge: The Construction of Knowledge in Artificial Intelligence’. *Social Studies of Science* 23, no. 3 (1 August 1993): 445–77. https://doi.org/10.1177/030631279302300302.

Friedman, Batya, and Helen Nissenbaum. 'Bias in Computer Systems'. *ACM Transactions on Information Systems* 14, no. 3 (1 July 1996): 330–47. https://doi.org/10.1145/230538.230561.

Gestel, Rob van, and Hans-Wolfgang Micklitz. 'Why Methods Matter in European Legal Scholarship: Methods in European Legal Scholarship'. *European Law Journal* 20, no. 3 (May 2014): 292–316. https://doi.org/10.1111/eulj.12049.

Glass, David J., and Ned Hall. 'A Brief History of the Hypothesis'. *Cell* 134, no. 3 (August 2008): 378–81. https://doi.org/10.1016/j.cell.2008.07.033.

Goodman, Bryce, and Seth Flaxman. 'European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"'. *AI Magazine* 38, no. 3 (2 October 2017): 50–57. https://doi.org/10.1609/aimag.v38i3.2741.

Greenberg, Andy. 'An AI That Reads Privacy Policies So That You Don't Have To'. *Wired*, 9 February 2018. https://www.wired.com/story/polisis-ai-reads-privacy-policies-so-you-dont-have-to/.

Hacker, Philipp. 'Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU Law'. *Common Market Law Review* 55, no. 4 (1 August 2018): 1143–85.

Hao, Karen. 'Can You Make an AI That Isn't Ableist?' MIT Technology Review. Accessed 29 November 2018. https://www.technologyreview.com/s/612489/can-you-make-an-ai-that-isnt-ableist/.

Hirsch, Reece, Kristin Hadgis, Morgan Lewis, and Bockius LLP. 'California's New, GDPR-Like Privacy Law Is A Game-Changer'. Bloomberg Law, 11 July 2018. https://news.bloomberglaw.com/privacy-and-data-security/insight-californias-new-gdpr-like-privacy-law-is-a-game-changer.

Ioannidis, John P. A. 'Why Most Published Research Findings Are False'. *PLOS Medicine* 2, no. 8 (30 August 2005): 696–701. https://doi.org/10.1371/journal.pmed.0020124.

Jensen, David. 'Data Snooping, Dredging and Fishing: The Dark Side of Data Mining A SIGKDD99 Panel Report'. *SIGKDD Explorations* 1, no. 2 (January 2000): 52–54.

Kamarinou, Dimitra, Christopher Millard, and Jatinder Singh. 'Machine Learning with Personal Data'. Legal Studies Research Paper 247/2016. Queen Mary University of London, School of Law, 7 November 2016. https://ssrn.com/abstract=2865811.

Kim, Tae Wan, Thomas Donaldson, and John Hooker. 'Mimetic vs Anchored Value Alignment in Artificial Intelligence'. *ArXiv:1810.11116 [Cs]*, 25 October 2018. http://arxiv.org/abs/1810.11116.

Kim, Tae Wan, and Bryan Routledge. 'Algorithmic Transparency, A Right To Explanation and Trust', June 2017, 31.

Kitchin, Rob. 'Big Data, New Epistemologies and Paradigm Shifts'. *Big Data & Society* 1, no. 1 (10 July 2014): 1–12. https://doi.org/10.1177/2053951714528481.

———. 'Thinking Critically about and Researching Algorithms'. *Information, Communication & Society* 20, no. 1 (2 January 2017): 14–29. https://doi.org/10.1080/1369118X.2016.1154087.

Koniaris, Marios, Ioannis Anagnostopoulos, and Yannis Vassiliou. 'Network Analysis in the Legal Domain: A Complex Model for European Union Legal Sources'. *Journal of*

Complex Networks 6, no. 2 (1 April 2018): 243–68. https://doi.org/10.1093/comnet/cnx029.

Kuner, Christopher. 'The European Commission's Proposed Data Protection Regulation: A Copernican Revolution in European Data Protection Law'. Bloomberg BNA Privacy and Security Law Report. Rochester, NY, 6 February 2012. https://papers.ssrn.com/abstract=2162781.

Leavy, Susan. 'Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning'. In *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering*, 14–16. GE '18. New York, USA: ACM, 2018. https://doi.org/10.1145/3195570.3195580.

Lecher, Colin. 'A Healthcare Algorithm Started Cutting Care, and No One Knew Why'. The Verge, 21 March 2018. https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 'Deep Learning'. *Nature* 521, no. 7553 (May 2015): 436–44. https://doi.org/10.1038/nature14539.

Lehman-Wilzig, Sam N. 'Frankenstein Unbound'. *Futures* 13, no. 6 (December 1981): 442–57. https://doi.org/10.1016/0016-3287(81)90100-2.

Lerman, Jonas. 'Big Data and Its Exclusions'. *Stanford Law Review Online* 66 (3 September 2013): 55–63. https://doi.org/10.2139/ssrn.2293765.

Levendowski, Amanda. 'How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem'. *Washington Law Review* 93 (24 July 2017): 579–630.

Mantelero, Alessandro. 'AI and Big Data: A Blueprint for a Human Rights, Social and Ethical Impact Assessment'. *Computer Law & Security Review* 34, no. 4 (1 August 2018): 754–72. https://doi.org/10.1016/j.clsr.2018.05.017.

Martin, Kirsten. 'Ethical Implications and Accountability of Algorithms'. *Journal of Business Ethics*, 7 June 2018. https://doi.org/10.1007/s10551-018-3921-3.

Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 'The Ethics of Algorithms: Mapping the Debate'. *Big Data & Society* 3, no. 2 (December 2016): 1-21. https://doi.org/10.1177/2053951716679679.

Mittelstadt, Brent Daniel, and Luciano Floridi. 'The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts'. *Science and Engineering Ethics* 22, no. 2 (April 2016): 303–41. https://doi.org/10.1007/s11948-015-9652-2.

Nothwang, William D., Michael J. McCourt, Ryan M. Robinson, Samuel A. Burden, and J. Willard Curtis. 'The Human Should Be Part of the Control Loop?' In *2016 Resilience Week (RWS)*, 214–20. Chicago, USA: IEEE, 2016. https://doi.org/10.1109/RWEEK.2016.7573336.

Nunn, Kenneth. 'Race, Crime and the Pool of Surplus Criminality: Or Why the "War on Drugs" Was a "War on Blacks"'. *UF Law Faculty Publications*, 1 January 2002. https://scholarship.law.ufl.edu/facultypub/107.

Pedreshi, Dino, Salvatore Ruggieri, and Franco Turini. 'Discrimination-Aware Data Mining'. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 560. Las Vegas, USA: ACM Press, 2008. https://doi.org/10.1145/1401890.1401959.

Picheta, Rob. ‘Passengers to Face AI Lie Detectors at EU Airports’. CNN Travel, 1 November 2018. https://www.cnn.com/travel/article/ai-lie-detector-eu-airports-scli-intl/index.html.

Searle, John R. ‘Is the Brain’s Mind a Computer Program?’ *Scientific American* 262, no. 1 (January 1990): 26–31. https://doi.org/10.1038/scientificamerican0190-26.

———. ‘Minds, Brains, and Programs’. *Behavioral and Brain Sciences* 3, no. 03 (September 1980): 417. https://doi.org/10.1017/S0140525X00005756.

Selbst, Andrew D., danah boyd, Sorelle Friedler, Suresh Venkatasubramanian, and Janet Vertesi. ‘Fairness and Abstraction in Sociotechnical Systems’. In *ACM Conference on Fairness, Accountability, and Transparency*, Vol. 1. 1. Rochester, USA, 2018. https://papers.ssrn.com/abstract=3265913.

Simonite, Tom. ‘When It Comes to Gorillas, Google Photos Remains Blind’. *Wired*, 11 January 2018. https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/.

Sweeney, Latanya. ‘Discrimination in Online Ad Delivery’. *SSRN Electronic Journal*, 28 January 2013. https://doi.org/10.2139/ssrn.2208240.

Talbot, David. ‘Amazon’s Same-Day Delivery Service Reinforces Inequality’. MIT Technology Review. Accessed 11 November 2018. https://www.technologyreview.com/s/601328/amazon-prime-or-amazon-redline/.

Tatman, Rachael. ‘Google’s Speech Recognition Has a Gender Bias’, 12 July 2016. https://makingnoiseandhearingthings.com/2016/07/12/googles-speech-recognition-has-a-gender-bias/.

Tonkens, Ryan. ‘Out of Character: On the Creation of Virtuous Machines’. *Ethics and Information Technology* 14, no. 2 (June 2012): 137–49. https://doi.org/10.1007/s10676-012-9290-1.

Trewin, Shari. ‘AI Fairness for People with Disabilities: Point of View’. IBM, 26 November 2018. http://arxiv.org/abs/1811.10670.

Turing, A. M. ‘Computing Machinery and Intelligence’. *Mind* LIX, no. 236 (1950): 433–60. https://doi.org/10.1093/mind/LIX.236.433.

Turkle, Sherry, and Seymour Papert. ‘Epistemological Pluralism: Styles and Voices within the Computer Culture’. *Signs* 16, no. 1 (1990): 128–57.

Vanian, Jonathan. ‘Unmasking A.I.’s Bias Problem’. Fortune, 25 June 2018. http://fortune.com/longform/ai-bias-problem/.

Vasconcelos, Marisa, Carlos Cardonha, and Bernardo Gonçalves. ‘Modeling Epistemological Principles for Bias Mitigation in AI Systems: An Illustration in Hiring Decisions’. *ArXiv:1711.07111 [Cs]*, 19 November 2017. https://doi.org/10.1145/3278721.3278751.

Wachter, Sandra, and Brent Mittelstadt. ‘A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI’. *Columbia Business Law Review* Forthcoming (13 September 2018). https://papers.ssrn.com/abstract=3248829.

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. ‘Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation’. *International Data Privacy Law* 7, no. 2 (3 June 2017): 76–99. https://doi.org/10.1093/idpl/ipx005.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 'Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR'. *Harvard Journal of Law & Technology* 31, no. 2 (2 November 2017). http://arxiv.org/abs/1711.00399.

Wajcman, Judy. 'Gender and Technology'. In *International Encyclopedia of the Social & Behavioral Sciences*, edited by Neil J. Smelser and Paul B. Baltes, 1st ed., 9:5976–5979. Amsterdam; New York: Elsevier, 2001.

Wallach, Wendell, and Colin Allen. *Moral Machines*. Oxford University Press, 2009. https://doi.org/10.1093/acprof:oso/9780195374049.001.0001.

Wenar, Leif. 'John Rawls'. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2017. Metaphysics Research Lab, Stanford University, 2017. https://plato.stanford.edu/archives/spr2017/entries/rawls/.

Williams, Brooks, and Shmargad. 'How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications'. *Journal of Information Policy* 8 (2018): 78–115. https://doi.org/10.5325/jinfopoli.8.2018.0078.

Willick, Marshall S. 'Artificial Intelligence: Some Legal Approaches and Implications'. *AI Magazine* 4, no. 2 (15 June 1983): 5–16. https://doi.org/10.1609/aimag.v4i2.392.

Zadrozny, Bianca. 'Learning and Evaluating Classifiers under Sample Selection Bias'. In *Twenty-First International Conference on Machine Learning*, 114. Alberta, Canada: ACM Press, 2004. https://doi.org/10.1145/1015330.1015425.

Zarsky, Tal. 'The Privacy–Innovation Conundrum'. *Lewis & Clark Law Review* 19, no. 1 (20 April 2015): 115–68.

———. 'The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making'. *Science, Technology, & Human Values* 41, no. 1 (January 2016): 118–32. https://doi.org/10.1177/0162243915605575.

———. 'Transparent Predictions'. *University of Illinois Law Review* 2013, no. 4 (10 September 2013): 1503–70.

Zerilli, John, Alistair Knott, James Maclaurin, and Colin Gavaghan. 'Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?' *Philosophy & Technology*, 5 September 2018, 1–23. https://doi.org/10.1007/s13347-018-0330-6.

Other publications

Baum, Seth D. 'A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy'. *Global Catastrophic Risk Institute Working Paper 17-1*, 2017. https://doi.org/10.2139/ssrn.3070741.

Information Commissioner's Office. 'Big Data, Artificial Intelligence, Machine Learning and Data Protection', 1 March 2017. https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf.

———. 'Data Protection Impact Assessments (DPIAs)', 14 May 2018. https://ico.org.uk/media/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/data-protection-impact-assessments-dpias-1-0.pdf.

Koene, Ansgar. 'Presentation of "a Governance Framework for Algorithmic Accountability and Transparency" at the European Parliament'. UnBias, 6 November 2018.

https://unbias.wp.horizon.ac.uk/2018/11/06/presentation-of-a-governance-framework-for-algorithmic-accountability-and-transparency-at-the-european-parliament/.

Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker. 'Algorithmic Impact Assessment: A Practical Framework for Public Agency Accountability'. AI Now, April 2018. https://ainowinstitute.org/aiareport2018.pdf.

'Rights Related to Automated Decision Making Including Profiling', 6 August 2018. https://icoumbraco.azurewebsites.net/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/rights-related-to-automated-decision-making-including-profiling/.

Science and Technology Committee. 'Robotics and Artificial Intelligence'. UK: House of Commons, 10 December 2016. https://publications.parliament.uk/pa/cm201617/cmselect/cmsctech/896/896.pdf.

'The Toronto Declaration: Protecting the Right to Equality and Non-Discrimination in Machine Learning Systems', 16 May 2018. https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf.

Internet resources

'AJL -Algorithmic Justice League'. AJL. Accessed 11 November 2018. https://www.ajlunited.org/.

'Definition of Bias'. Oxford Dictionaries | English. Accessed 7 November 2018. https://en.oxforddictionaries.com/definition/bias.

'Definition of Data'. Oxford Dictionaries | English. Accessed 12 November 2018. https://en.oxforddictionaries.com/definition/data.

'Fairness, Accountability, and Transparency in Machine Learning'. Accessed 11 November 2018. http://www.fatml.org/.

'Part-Time Employment Rate'. OECD. Accessed 16 November 2018. https://doi.org/10.1787/f2ad596c-en.

'Principles for Accountable Algorithms and a Social Impact Statement for Algorithms: FAT ML'. Accessed 12 November 2018. http://www.fatml.org/resources/principles-for-accountable-algorithms.

'Project Overview Algorithmic Justice League'. MIT Media Lab. Accessed 11 November 2018. https://www.media.mit.edu/projects/algorithmic-justice-league/overview/.

'Temporary Employment'. OECD. Accessed 16 November 2018. https://doi.org/10.1787/75589b8a-en.