DOI: 10.14515/monitoring.2021.1.1911

A. Karatzogianni

RESEARCH DESIGN FOR AN INTEGRATED ARTIFICIAL INTELLIGENCE ETHICAL FRAMEWORK

For citation:
Karatzogianni A. (2021) Research Design for an Integrated Artificial Intelligence Ethical Framework. *Monitoring of Public Opinion: Economic and Social Changes.* No. 1. P. 31–45. https://doi.org/10.14515/monitoring.2021.1.1911.
Правильная ссылка на статью:
Караджоянни А. Интеграция этических оснований искусственного интеллекта: план исследования // Мониторинг общественного мнения: экономические и социальные перемены. 2021. № 1. С. 31—45. https://doi.org/10.14515/monitoring.2021.1.1911. (In Eng.)

DOI: 10.14515/monitoring.2021.1.1911

RESEARCH DESIGN FOR AN INTEGRATED ARTIFICIAL INTELLIGENCE ETHICAL FRAMEWORK

Athina KARATZOGIANNI[1] *— Professor in Media and Communication*
E-MAIL: athina.k@le.ac.uk
https://orcid.org/0000-0002-6161-4423

[1] University of Leicester, Leicester, England

Abstract. Artificial Intelligence (AI) regulatory and other governance mechanisms have only started to emerge and consolidate. Therefore, AI regulation, legislation, frameworks, and guidelines are presently fragmented, isolated, or co-exist in an opaque space between national governments, international bodies, corporations, practitioners, think-tanks, and civil society organisations. This article proposes a research design set up to address this problem by directly collaborating with targeted actors to identify principles for AI that are trustworthy, accountable, safe, fair, non-discriminatory, and which puts human rights and the social good at the centre of its approach. It proposes 21 interlinked substudies, focusing on the ethical judgements, empirical statements, and practical guidelines, which manufacture ethicopolitical visions and AI policies across four domains: seven tech corporations, seven governments, seven civil society actors, together with the analysis of online public debates. The proposed research design uses multiple research techniques: extensive mapping and studies of AI ethics policy documents and 120 interviews of key individuals, as well as assorted analyses of public feedback discussion loops on AI, employing digital methods on online communities specialising in AI debates. It considers

ИНТЕГРАЦИЯ ЭТИЧЕСКИХ ОСНОВАНИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА: ПЛАН ИССЛЕДОВАНИЯ

КАРАДЖОЯННИ Афина — профессор, департамент медиа и коммуникации, Лестерский университет, Лестер, Англия
E-MAIL: athina.k@le.ac.uk
https://orcid.org/0000-0002-6161-4423

Аннотация. Сегодня механизмы управления и регулирования для искусственного интеллекта (ИИ) только начинают формироваться. Принципы регулирования, правовые основы и «дорожные карты» развития ИИ фрагментарны и разрозненны, они существуют в теневом пространстве между национальными государствами, международными институциями, бизнес-корпорациями, сообществами практиков, аналитическими центрами и организациями гражданского общества. В статье предлагается план исследования, направленного на решение данной проблемы и предполагающего сотрудничество с ключевыми акторами с целью определения принципов развития ИИ, которые были бы надежными, понятными, безопасными, справедливыми, беспристрастными, ставили бы в центр права человека и общественное благо. Дизайн исследовательского проекта предполагает проведение взаимосвязанных мини-исследований (21 шт.), направленных на анализ этических суждений, эмпирических фактов и практических рекомендаций, формирующих этику и политику в отношении ИИ. План предполагает проведение исследований в семи технологических корпорациях, в административных органах семи национальных государств и в семи ор-

DOI: 10.14515/monitoring.2021.1.1911

novel conceptual interactions communicated across the globe, expands the regulatory, ethics, and technological foresight, both at the individual level (autonomy, identity, dignity, privacy, and data protection) and the societal level (fairness/equality, responsibility, accountability and transparency, surveillance/datafication, democracy and trust, collective humanity and the common good). By producing an innovative, intercontinental, multidisciplinary research design for an Ethical AI Standard, this article offers a concrete plan to search for the Holy Grail of Artificial Intelligence: Its Ethics.

Keywords: artificial intelligence, artificial intelligence regulation, ethical artificial intelligence standard, artificial intelligence policy, multidisciplinary research design, artificial intelligenece ethics

ганизациях гражданского общества, а также анализ публичных дебатов в интернете. Дизайн исследования включает несколько методов: подробное картирование и изучение политических и юридических документов, касающихся этики ИИ; 120 экспертных интервью; анализ циклов общественного обсуждения ИИ в специализированных онлайн-сообществах. Исследование направлено на анализ новых концептуальных взаимодействий между участниками процесса по всему миру, а также на расширение возможностей нормативного, этического и технологического прогнозирования как на индивидуальном уровне (вопросы автономии, идентичности, достоинства, конфиденциальности и защиты данных), так и на уровне общества (справедливость/равенство, ответственность, подотчетность и прозрачность, надзор/датафикация, демократия и доверие, общественный гуманизм и общее благо). Представляя дизайн инновационного, международного и междисциплинарного исследования этического стандарта ИИ, статья предлагает конкретный план поиска Святого Грааля искусственного интеллекта — его этических оснований.

Ключевые слова: искусственный интеллект, регулирование искусственного интеллекта, этический стандарт искусственного интеллекта, политика в отношении искусственного интеллекта, междисциплинарный исследовательский дизайн, этические основания искусственного интеллекта

DOI: 10.14515/monitoring.2021.1.1911

Introduction

In Bristol, Artificial Intelligence (AI) has given us a 'youth score' computer programme, which combines crime data, housing information, and links them to others viewed as high-risk, together with information about the youth's parents and domestic incidents. It also feeds school attendance records in. The police and social workers then surge resources towards high-risk cases and away from those that do not meet the indicators. In Philadelphia, face-to-face interviews with parole officers have been overtaken by predictive algorithms to set probation rules. In Amsterdam, algorithms identify welfare fraud risks and allocate credit. In February 2020, *The New York Times* observed that we have already entered an era when an algorithm grants freedom or takes it away[1].

Daily, advanced democratic societies are forced into operating more digitally by default. In this context, our social lives are increasingly governed by algorithms. AI software predicts who will commit a crime and making probation decisions, which demographics can have loans, who to provide healthcare to, who to hire or admit to university, even guiding sentences handed down by judges. Black box and unaccountable technologies offered by unregulated private companies have a profound effect on authority, trust, and transparency, with profound consequences for justice, education, and welfare in societies around the world.

And yet, there is a lack of a global ethical agreement on Artificial Intelligence (AI), although it poses the most significant moral challenge of our time. We are remarkably short of evidence-based social science research on how these systems are working now, how they are governed, and mainly how ethical standards are being practically applied, especially regarding social and economic inclusion [Jobin, Ienca, Vayena, 2019; Redden, Dencik, Warne, 2020; Sanchez-Monedero, Dencik, Edwards, 2020]. Because this lack represents a severe test of humane values, it drives the central vision of this research design experiment: to propose a research design to develop an innovative, intercontinental, multidisciplinary integrated framework for an Ethical AI Standard. The most innovative aspect of this research design is a targeted programme to select, analyse, cross-examine, integrate and expand inputs and debates from twenty-one tech corporations, government organisations, civil society actors, and the analysis of debates generated on social media platforms by the general public, globally. This can be achieved by investigating in-depth ethicopolitical judgements, empirical statements, and practical guidelines produced in public AI policy documents, interviews with experts and practitioners, and debates circulating in the digital public domain. To create and implement ethical and legitimate AI governance, stakeholders need to be confronted with their own and others' ethicopolitical visions and discourses. They must also be confident that the researchers understand the practicalities of delivering advanced AI technology and the concerns of individuals and organisations requiring privacy and transparency in government and corporate policy in this area. The overall objective is to investigate the ethical and political visions of corporate, governmental, and civil society organisations, and the general public and cross-examine these with the direct engagement of interview participants.

[1] Metz C., Satariano A. (2020) An Algorithm that Grants Freedom, or Takes It Away. *The New York Times*. Feb. 6. URL: https://www.nytimes.com/2020/02/06/technology/predictive-algorithms-crime.html (accessed: 27.02.2021).

DOI: 10.14515/monitoring.2021.1.1911

With global, cross-sector, specialised, and general population input, this research design is set up to produce an integrated framework for Ethical Artificial Intelligence. Such a quest is the Holy Grail of technology ethics because of the high stakes involved in the use and abuse of Artificial Intelligence, which has critical consequences for humanity's future [Bostrom, 2014; Floridi, 2015; Harari, 2016]. The rapid development of AI and its application in fields as diverse as medical surgery, autonomous cars, and military robots, together with all-purpose use simulations of machine learning, has caused growing concerns about the unknown impact of AI in an anarchic world characterised by secretive commercial and nation-state competition [Kaplan, 2015; Acemoglu, Restrepo, 2018]. Artificial intelligent machines are advanced software systems. The questions are who designs and is in charge of these systems, who controls, regulates, and can have the data to intervene in time when AI is not serving the purpose with which it was designed [Garfinkel, Matthews, Shapiro, Smith, 2017]. Human values must be able to shape this future, and this future has to include everyone. By searching for AI, humanity is also searching for the best future for a human species capable of governing AI and developing an AI that displays the emotional and social intelligence to work with humans. Above all, we need an AI that compensates for rather than exploits human limitations because it understands blind spots in human cognition, memory, judgement, and attention, even empathy [van Dijk, 2014].

There are already remarkable visions of training AI to predict how a human would punish AI, when it ethically deviates. With rapid advancements in natural language translation, voice recognition, and a massive amount of computational time and space, whereby AI breathes, as it trains on human text, humans are often confused, feeling that the AI is human. Accordingly, there is concern that AI will bring a sense of loss: the uniqueness of being human. When this AI comes forward to interact with humans, it must go ahead with human values. We need to consider what value re-alignment is required in this partnership [Floridi, 2018]. What should humanity want from AI's future, in an era when machines will change human behaviour as never before?[2]. Although humanity still understands little about how children are learning and have made little progress on the workings of human consciousness, there is nevertheless a pervasive use of AI that is unregulated, under little control and confronts legislation that is too slow for the accelerated sped up pace with which AI is evolving[3]. The symptoms of this unbounded acceleration are already in plain sight:

— Fake news and disinformation architectures which pose risks of populism, radicalism, violent extremism together with algorithmic interference[4] [Sumpter, 2018];
— Gender, race, class, and other algorithmic bias [O'Neil, 2016; Chouldechova, 2017];
— Emerging issues in employment, health, education;

[2] Guillén M., Reddy S. (2018) We Know Ethics Should Inform AI. But Which Ethics? *World Economic Forum*. 26 July. URL: https://www.weforum.org/agenda/2018/07/we-know-ethics-should-inform-ai-but-which-ethics-robotics/ (accessed: 27.02.2021).

[3] The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. (2018) URL: https://www.repository.cam.ac.uk/bitstream/handle/1810/275332/1802.07228.pdf?sequence=1 (accessed: 27.02.2021).

[4] Ong J. C., Cabañes J. V. A. (2017) Architects of Networked Disinformation: Behind the Scenes of Troll Accounts and Fake News Production in the Philippines (Public Report). URL: https://newtontechfordev.com/wp-content/uploads/2018/02/ARCHITECTS-OF-NETWORKED-DISINFORMATION-FULL-REPORT.pdf (accessed: 27.02.2021).

DOI: 10.14515/monitoring.2021.1.1911

— Future of work, quantification, recruitment bias, digital labour and gig economy [Codagnone, Karatzogianni, Matthews, 2018];
— Data justice, whistle-blowing, legal studies, digital rights, data inequality[5] [Hintz, Dencik, Wahl-Jorgensen, 2018]; and
— Superintelligence what Bostrom, Dafoe and Flynn[6] call 'mind crime prevention', ensuring that advanced AI is governed in such a way that maltreatment of sentient digital minds is avoided or minimized.

Equally, we may see future potential resistance to Artificially Intelligent machines, which would predictably see future AI-resisting social movements and non-state actors taking digital activism and cyberconflict to unimaginable new heights.

AI is now a top research priority. In the past few years, there has been a proliferation of reports on AI from governmental and other organisations[7]. Let us consider the two most recent European commission responses to critical issues arising from AI with the publication of *Artificial Intelligence: A European Perspective*[8], *A Draft Ethics Guidelines for Trustworthy AI*[9]; *A Definition of AI: Main Capabilities and Disciplines*[10].

[5] César J., Debussche J., van Asbroeck B. (2017) White Paper — Data Ownership in the Context of the European Data Economy: Proposal for a New Right. *Bird & Bird*. February. URL: https://www.twobirds.com/en/news/articles/2017/global/data-ownership-in-the-context-of-the-european-data-economy (accessed: 27.02.2021).

[6] Bostrom N., Dafoe A., Flynn C. (2018) Public Policy for Superintelligent AI: A Vector Field Approach. URL: https://nickbostrom.com/papers/aipolicy.pdf (accessed: 27.02.2021).

[7] Executive Office of the President National Science and Technology Council Committee on Technology (2016) Preparing for the Future of Artificial Intelligence. October. URL: https://info.publicintelligence.net/WhiteHouse-ArtificialIntelligencePreparations.pdf (accessed: 26.02.2021); UK Government Office for Science (2015) Artificial Intelligence: Opportunities and Implications for the Future of Decision Making. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/566075/gs-16-19-artificial-intelligence-ai-report.pdf (accessed: 26.02.2021); UK House of Commons Science and Technology Committee (2016) Robotics and Artificial Intelligence. Fifth Report of Session 2016—17. URL: https://publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/145.pdf (accessed: 26.02.2021); European Economic and Social Committee (2017) Artificial Intelligence — The Consequences of Artificial Intelligence on the (Digital) Single Market, Production, Consumption, Employment and Society (Own-Initiative Opinion). URL: https://www.eesc.europa.eu/en/our-work/opinions-information-reports/opinions/artificial-intelligence (accessed: 27.02.2021); European Parliament Policy Department (2016) European Civil Law Rules in Robotics. URL: EUROPEAN CIVIL LAW RULES IN ROBOTICS (europa.eu) (accessed: 27.02.2021); Council of Europe Committee of experts on internet intermediaries (2017) Study on the Human Rights Dimensions of Automated Data Processing Techniques (in Particular Algorithms) and Possible Regulatory Implications. URL: https://rm.coe.int/study-hr-dimension-of-automated-data-processing-incl-algorithms/168075b94a (accessed: 27.02.2021); Ministry of Economic Affairs and Employment of Finland (2017) Finland's Age of Artificial Intelligence. Turning Finland into a leading country in the application of artificial intelligence. Objective and recommendations for measures. URL: http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/160391/TEMrap_47_2017_verkkojulkaisu.pdf?sequence=1&isAllowed=y (accessed: 27.02.2021); France Intelligence Artificielle (2017) Rapport de Synthèse — France IA. URL: https://www.economie.gouv.fr/files/files/PDF/2017/Rapport_synthese_France_IA_.pdf (accessed: 27.02.2021); Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions (2018) Artificial Intelligence for Europe. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN (accessed: 27.02.2021); European Commission's High-Level Expert Group on Artificial Intelligence (2018) Draft Ethics Guidelines for Trustworthy AI. 18 December. URL: Draft Ethics guidelines for trustworthy AI | Shaping Europe's digital future (europa.eu) (accessed: 27.02.2021); European Group on Ethics in Science and New Technologies (2018) Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems. URL: https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf (accessed: 27.02.2021); Deloitte (2017) Study on Emerging Issues of Data Ownership, Interoperability, (Re)-Usability and Access to Data, and Liability. URL: http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=51486 (accessed: 27.02.2021).

[8] European Commission Science Hub (2018) Artificial Intelligence: A European Perspective. URL: https://ec.europa.eu/jrc/en/publication/artificial-intelligence-european-perspective (accessed: 27.02.2021).

[9] European Commission's High-Level Expert Group on Artificial Intelligence (2018) Draft Ethics Guidelines for Trustworthy AI. 18 December. URL: Draft Ethics guidelines for trustworthy AI | Shaping Europe's digital future (europa.eu) (accessed: 27.02.2021).

[10] European Commission's High-Level Expert Group on Artificial Intelligence (2019) A Definition of AI: Main Capabilities and Disciplines. URL: https://www.aepd.es/sites/default/files/2019-12/ai-definition.pdf (accessed: 27.02.2021).

DOI: 10.14515/monitoring.2021.1.1911

The AI European perspective report differentiates between individual and collective implications of AI: autonomy, identity, dignity, privacy, and data protection at the individual level. Further, it recognises that AI dramatically affects the societal level: fairness and equity, responsibility, accountability and transparency, privacy in terms of surveillance/datafication, democracy and trust, and collective identity and good life. In reviewing key ethical and social issues in AI, it identifies two new rights:

(1) The right to meaningful human contact, whereby every person may refuse to be cared for by a robot, and robots should respect humans' autonomy in decision-making;
(2) The right to refuse being profiled, tracked, measured, analysed, coached, or manipulated.

The authors also prescribe responsible AI design, which engages critically with civil society, establishes multi-stakeholder fora to promote such public debate translating outcomes to strategies for AI enforcing ethical and social values, and the design practice to address potential sources of the AI system from selection of team, to labelling and training data evaluation of outputs, and assessments of outcomes [11]. Given the significance of these findings, it is not unreasonable to anticipate these issues making up additional protocols to the European Convention of Human Rights in the near future.

Where are we with AI ethical governance at present? Arguably we are at an elementary stage, and this area of research requires urgent advancement. Accordingly, the 'Draft Ethics Guidelines for Trustworthy AI' aimed at maximising the benefits of AI while minimising its risks, advocating that a human-centric approach to AI is needed to create 'Trustworthy AI', respecting fundamental human rights, ensuring an 'ethical purpose', and asserting that it should be technically robust. The Guidelines operationalise the requirements of ethical purpose and robustness, providing us with a concrete assessment list. This list includes the incorporation of Trustworthy AI from the earliest design phase: accountability, data governance, design for all, governance of AI autonomy (human oversight), non-discrimination, respect for human autonomy, respect for privacy, robustness, safety, and transparency. The key guidance embraces technical and non-technical methods to implement while keeping in mind ethical considerations when recruiting the team building the system, the system itself, the testing environment, and potential applications. Additionally, stakeholders (customers, employees) should have transparent and proactive information regarding the AI's capabilities and limitations, ensuring traceability.

Suppose we want to promote ethical AI policies and practices AI to the level of strategic goals for leading organisations. In that case, AI should be part of the organisation's culture, embedded in deontological chapters or codes of conduct, ensuring stakeholders' inclusion in the AI development and diversity in the team producing it. This enables us to foresee training and education in Trustworthy AI and ensure a specific process for accountability governance [12]. Assessing Trustworthy AI includes

[11] European Commission Science Hub (2018) Artificial Intelligence: A European Perspective. URL: https://ec.europa.eu/jrc/en/publication/artificial-intelligence-european-perspective (accessed: 27.02.2021).

[12] European Commission's High-Level Expert Group on Artificial Intelligence (2018) Draft Ethics Guidelines for Trustworthy AI. 18 December. URL: Draft Ethics guidelines for trustworthy AI | Shaping Europe's digital future (europa.eu) (accessed: 27.02.2021).

DOI: 10.14515/monitoring.2021.1.1911

accountability, data governance, design for all, governing AI autonomy, non-discrimination, respect for privacy, robustness, reliability, reproducibility, and accuracy through data usage and control fall-back plan, safety, transparency, and traceability[13].

There are important choices to be made. Geopolitically, this approach differentiates European AI Ethics from the unsustainable and undemocratic development of AI involving massive surveillance and control of populations in the Middle East and North Africa, and East Asia. Also, in relation to oligopolistic algorithmic governance by tech companies without significant governmental regulatory commitment to democracy and trust in North America. The World Intellectual Property Organisation (WIPO) Technology Trends 2019 identifies deep learning as the fastest growing technique with an increase of 175 % between 2013 and 2016[14]. Crucially, deep learning is 'black box' AI, which relies on neural networks, in contrast to 'white box' AI, where all the code lines are explicit. Companies represent twenty six of the top thirty applicants, and it is striking that just four are university or public research organisation. IBM tops the list, followed by Microsoft, and out of the top twenty, twelve are based in Japan, three are from the US, and two are from China. The report identifies the geographical origin of the university and public research organisations in the top 500 as China, US, Korea, Taiwan, Europe, Japan, Russian Federation, Saudi Arabia, in that order with China clearly by far dominating patents in that field[15].

A Global Ethical Problem: In Europe, Ethics and AI involve significant ethical judgements, empirical statements, and practical guidelines, which rely heavily on the direct adaptation of what we could call 'social-democratic humanism'. To the West, the United States is a prime example of drawing from a 'neo-liberal humanism', whereby the individual knows best, the customer will decide, and the company will create a product which the customer does not even know that they already want. Here, dataism and techno-utopianism are the resulting principles. Humans will accept all, as long as they can stay in the data flow and take advantages of AI leaps. Those particular humans that can enhance their body and life with AI will evolve as a new elite of superhumans, and those that are rendered obsolete and useless by AI will be left behind and outevolved. The era of the masses is over. These alarming Darwinist ethics are drawing from 'evolutionary humanism' ideologies of the past (eugenics is a prime example). Harari [2016] inspired this line of argumentation in *Homo Deus,* where he traces some continuities and discontinuities in the AI ethicopolitical visions. These have become the philosophical departure point of this project.

To integrate and produce an Ethical AI Standard, this research design is set up to answer the following key research question: What are the competing AI ethicopolitical visions of key actors in the field of AI?

Designing for AI Ethics Research

With this research question in mind, this research design experiment is set up to analyse, trace, evaluate, select, integrate and expand diverse and fragmented ethico-

[13] Ibidem.

[14] World Intellectual Property Organisation (2019) WIPO Technology Trends 2019 — Artificial Intelligence. URL: https://www.wipo.int/publications/en/details.jsp?id=4386 (accessed: 27.02.2021).

[15] Ibid.: 32.

DOI: 10.14515/monitoring.2021.1.1911

political visions of AI, considering the proposals envisaged by the European AI ethicists above. The project utilises directly European recommendations because they are crucial for any Ethical AI Standard. Still, it does so by engaging multistakeholder fora by interviewing key actors, engaging with civil society, and promoting public debate beyond the European Union countries. Furthermore, it discusses the European AI ethical framework with key stakeholders, governments, corporations, civil society actors, and the global public. It anticipates an open and reflexive critique that will take a potential project forward. The research design offered here relies on the following research techniques and objectives:

Project Objective 1 (PO1): Maps key AI ethicopolitical frameworks in circulation by the 21 key actors. This will involve the collection of AI policy documents produced by three sets of key players: seven tech corporations (Google, Amazon, Facebook, Apple, Microsoft, Tesla, and Alibaba), seven governmental organisations (China, Japan, United States, European Union, Australia, India, and South Africa), seven civil society actors (The Partnership on AI, Open AI, Association for the Advancement of Artificial Intelligence, European Association for Artificial Intelligence, Future of Life Institute, Society for the Study of Artificial Intelligence and the Simulation of Behaviour, and the Machine Intelligence Research Institute).

(PO2): Investigates ethicopolitical visions on AI across seven tech corporations. Examines ethicopolitical visions by seven tech corporations and juxtaposes these with the findings from PO1, asking interview participants to compare their views concerning PO1. The specific tech corporations (dubbed the internet oligopoly with the acronym GAFAM: Google, Amazon, Facebook, Apple, Microsoft) are chosen because they have been recently embattled in ethical issues publicly and extensively. Alibaba is selected because of the sheer scale of AI application involved in its trade. Tesla is examined because it has ranked as the world's best-selling plug-in passenger car manufacturer and works across several technological innovation domains applying AI. Crucially, Tesla founder Elon Musk has repeatedly advocated strong AI regulation in public.

The GAFAM tech corporations have been recently embattled in ethical issues publicly and extensively. Examples of why Google is chosen involves recent reports of the crowd workers outsourced to support a contract the company had with the US military on drones and the extensive ethical issues this brought up with employees with the company, resigning and demanding adherence to the company motto 'Do No Evil'. Google owns YouTube, which has also been controversial in terms of content moderation in relation to online radicalisation videos appearing next to advertisements, with companies and governments withdrawing advertising from the platform. Facebook has been embroiled in the Cambridge Analytica scandal, disinformation and potential impact during elections around the world. Subsequently, the slow and inadequate response the company rendered against its public critics, in relation to privacy, together with a ramification of its advertising practices, and alleged interference on its platform, potentially influencing the election and referendum results in several countries (e. g., the US 2016 Elections, Brexit), together with the ultimate 'hacking' of democratic institutions. The closer integration of WhatsApp and Instagram, which the company acquired, has caused widespread public criticism and an array of ethical issues relating to children and youth's use of their platforms in particular.

DOI: 10.14515/monitoring.2021.1.1911

Amazon, Apple, Microsoft are tech companies that have also involved much in the development of advanced software systems and are considered influential players in mobile and desktop applications and hardware. Alibaba and Tesla are included here for their significant record in this domain. Alibaba is chosen because of the sheer scale of AI applications involved in its trade. During 'Singles Day' Alibaba processed 325,000 orders per second through pop-up stores selling products fitted with Virtual Reality mirrors, using an AI fashion consultant matching items. One day, it sold 25 billion dollars' worth of goods[16]. Tesla is chosen for this specific reason: Tesla founder Elon Musk has repeatedly advocated strong AI regulation in public for the past decade. 'It needs to be a public body that has insight and then oversight to confirm that everyone is developing AI safely. This is extremely important. I think the danger of AI is much greater than the danger of nuclear warheads by a lot and nobody would suggest that we allow anyone to build nuclear warheads if they want. That would be insane'[17].

(PO3): Investigates ethicopolitical visions on AI across seven governments. Examines ethicopolitical visions by seven governmental organisations and asks interview participants to compare their views in relation to PO2. Seven governmental organisations are investigated (China, Japan, the United States; the European Union — focus on Germany France and the Nordic-Baltic Eight (NB8)[18]; Australia, India and South Africa). AI is ultra-nationalised and governments are pressured into the impossible position to develop AI policies that are competitive while protecting citizen rights (transparency, accountability, privacy, equal treatment, non-discrimination, mitigation of harmful impacts). In recent years, these actors have released AI visions. Interviews with policymakers in government will involve direct questions about the AI ethicopolitical visions expressed by actors in PO2.

Although the first AI patent filings were made in Japan in the 1980s, the field has been overtaken by China and the United States. Since 2014, China has been the leader in a number of first patents filed. In 2017, the State Council announced the 'Next Generation AI Development Plan' with the ambition of becoming the world's primary innovation centre by 2030, followed up by a 'Three Year Plan to Promote the Development of the New-Generation AI Industry'[19]. In the United States, three reports were released in 2016: 'Artificial Intelligence, Automation, and the Economy'; 'Preparing for the Future of Artificial Intelligence'; and 'The National Artificial Intelligence Research and Development Plan', while in 2018, a Select Committee on Artificial Intelligence was announced[20]. The European Union is going to be researched as an intergovernmental organisation, however, with the understanding the specific countries are going to be investigated in more depth, Germany, France, particularly the Nordic-Baltic Eight (NB8), because they made a joint statement in May 2018 to enhance

[16] European Commission Science Hub (2018) Artificial Intelligence: A European Perspective. P. 60. URL: https://ec.europa.eu/jrc/en/publication/artificial-intelligence-european-perspective (accessed: 27.02.2021).

[17] Young A. (2018) Musk says AI 'More Dangerous Than Nukes' — Expert Stays Optimistic *SecurityBrief.eu.* 13 March. URL: Musk says AI 'more dangerous than nukes'-expert stays optimistic (securitybrief.eu) (accessed: 27.02.2021).

[18] Denmark, Estonia, Finland, Iceland, Latvia, Lithuania, Norway, and Sweden.

[19] World Intellectual Property Organisation. (2019) WIPO Technology Trends 2019 — Artificial Intelligence. P. 127. URL: https://www.wipo.int/publications/en/details.jsp?id=4386 (accessed: 27.02.2021).

[20] Ibid.: 126.

DOI: 10.14515/monitoring.2021.1.1911

access for data for AI, stating that they want to 'avoid unnecessary regulation that could get in the way of this fast-developing field' [21].

Apart from these top players, Australia, India, and South Africa are chosen to provide a more intercontinental perspective. With headlines such as 'Australia lags on AI, automation' [22] and 'Australian needs to embrace automation or risk missing a 2.2 trillion-dollar boom' [23], Australia is a case worth studying further. India is chosen because of their #AIforAll approach. In their 'National Strategy for Artificial Intelligence #AIforAll', whereby '#AIforAll will focus on harnessing collaborations and partnerships, and aspires to ensure prosperity for all. Thus, #AIforAll means technology leadership in AI for achieving the greater good' [24]. In Africa, we will focus on South Africa, together with a broader interest in understanding AI for development and organisations such as Machine Intelligence Institute of Africa [25]. In addition, events such as AI for Good Global Summit and United Nations AI conference are the type of events where actors' interplay can be observed, and potential fieldwork interviews can be conducted.

(PO4): Investigates ethicopolitical visions on AI across seven AI-specialised civil society organisations. Examines ethicopolitical visions by seven AI-specialised civil society organisations and asks interview participants to compare their views concerning PO2 and PO3. The current sample includes the following organisations: The Partnership on AI, Open AI, Association for the Advancement of Artificial Intelligence, European Association for Artificial Intelligence, Future of Life Institute, Society for the Study of Artificial Intelligence and the Simulation of Behaviour, and the Machine Intelligence Research Institute. As PO1 kicks off mapping the AI policy environment and during fieldwork for PO2 and PO3, we will acquire further insights on which are the most relevant specialised organisations to investigate. The reason we are interested in civil society organisations specialised on AI and not generally, for example, in privacy, transparency, or digital rights organisations is because there are several areas of technical and policy expertise involved in this area, and we do require a sufficient level of specialisation to integrate insights and principles from.

(PO5): Investigates public receptions of ethicopolitical visions identified in PO1–PO4. Examines ethicopolitical visions of the 21 actors, and how they are received by the general population on social media platform debates. We will choose to collect public debates across Facebook (groups such as 'Artificial General Intelligence'; 'Artificial Gods', 'Real AGI', 'Artificial Intelligence and Deep Learning'), Twitter, Sina Weibo, as the most dominant globally, and online forums specialising on AI debates, such as 'The Artificial Intelligence Forum' [26], 'Ai dreams' [27], on 'Reddit' [28], 'Quora's Artificial General

[21] Ibid.: 127.

[22] Australia lags on AI, automation. (2019) *Erpinews*. October 24. URL: https://erpinnews.com/australia-lags-on-ai-automation (accessed: 27.02.2021).

[23] Dunn M. (2018) Australian needs to embrace automation or risk missing a 2.2 trillion boom. News.com.au. June 1. URL: https://www.news.com.au/technology/innovation/inventions/australia-needs-to-embrace-automation-or-risk-missing-a-22-trillion-boom/news-story/23b2608dec515e3749601d46bac6143d (accessed: 27.02.2021).

[24] Future of Life Institute. AI Policy — India. URL: https://futureoflife.org/ai-policy-india/?cn-reloaded=1 (accessed: 27.02.2021).

[25] URL: http://machineintelligenceafrica.org/ (accessed: 26.02.2021).

[26] For more details, see URL: https://ai-forum.com/ (accessed: 27.02.2021).

[27] For more details, see URL: https://aidreams.co.uk/ (accessed: 27.02.2021).

[28] For more details, see URL: https://www.reddit.com/r/machinelearning (accessed: 27.02.2021).

DOI: 10.14515/monitoring.2021.1.1911

Intelligence' [29]. The purpose is to conduct social network analysis and semantic analysis of what are the dominant actors, relations and debates in the digital public domain, what are the ethical judgements and empirical statements in circulation and particularly how the 21 actors we are interested in are received in those circles.

(PO6): Produces an Integrated framework for an Ethical AI Standard (iExIST). The final work package will first select, integrate and expand the AI policy mapping from PO1. It will establish and synthesise the themes and principles informing the 21 actors plus public circulation ethical judgements, juxtapose these to empirical statements accordingly (PO2, PO3, PO4, PO5), integrate best practical guidelines, and then disseminate these findings to the actors interviewed to create a feedback loop of best approaches to produce the final framework for an Ethical AI standard.

Research Techniques

AI policy document analysis for PO1, using NVIVO, will enable us to collect, organise and analyse content from interviews, social media data, YouTube videos and web pages. In this way, we can describe and document data in a highly organised fashion, which will help both during critical multimodal discourse analysis and when data are shared after the research projects end. Collection and analysis of primary (reports, documents, legislation, policy assessments) and secondary (academic/other) materials, focusing on process-tracing each stakeholder's role in the evolving system of AI implementation. We will concentrate on three departments in each country: typically ministries of health/social security, education, and justice. In some countries, AI responsibilities are more dispersed and include departments dealing with media/culture, technical assurance, policing, and security.

Semi-structured interviews for PO2, PO3, PO4: Research interviews will be conducted with the primary stakeholders and their attitudes and beliefs regarding AI ethical and social issues. The type of questions will include beliefs (what people believe to be the case); attitudes (what people would prefer to be the case); behaviour (examples from their own experience as practitioners, policymakers, and activists). Interviewee attributes will be recorded and anonymised when this is necessary during consent and ethical issues emerging. Interviews are essential in establishing what our key actors think about the changing nature of AI in society and their more general attitudes towards current practice and procedure. While it is possible to obtain some of this information from policy documents, our emphasis on the interview will allow us to draw out the respondents at length regarding their thoughts on real world issues. It will enable us to ask open-ended questions and permits the respondent to talk more freely. It is appropriate for a project in which we wish to gather rich ethnographic about working with AI.

We also emphasise interviews because of our desire to undertake a degree of the process- tracing. This is to establish the decisions and attitudes that underpinned existing protocols and responses to particular AI issues in the respective countries. Within this focused inquiry, we will be able to reconstruct specific practical episodes based on the interview testimony and then compare accounts to give us a sophisti-

[29] For more details, see URL: https://www.quora.com/What-is-artificial-general-intelligence-AGI (accessed: 27.02.2021).

DOI: 10.14515/monitoring.2021.1.1911

cated picture of particular regulatory phenomena. Expert interviews are especially appropriate to study regulation-building since they can illuminate hidden elements of social action that are not clear from analysis of political outcomes using documentary materials. As others have argued, the existing literature on AI is fragmented and predominantly focused on the formal, legal, and informational rather than social aspects of regulation. Our emphasis on semi-structured interviews distinguishes between formal and informal processes and seeks to unpick some everyday activity around AI implementation. This approach will also encourage the co-production of knowledge during the project and beyond.

Social Network Analysis (SNA) and semantic analysis for PO5–PO6: The examination of transnational debates surrounding the 21 actors, and debates on AI ethics relations to the offline world. Wasserman and Faust [1994] explain that SNA 'provides a precise way to define important social concepts, a theoretical alternative to the assumption of independent social actors, and a framework for testing theories about structured social relationships' [ibid.: 17]. SNA is appropriate for the analysis and the investigation of 'kinship patterns, community structure, interlocking directorships and so forth' [Scott, 2000: 2]. SNA supports examining different social entities or social units, including individual, corporate, or collective social units [Wasserman, Faust, 1994: 16—21]. Key social media platforms, such as Facebook, will be examined to understand the formation of online networks and coalitions, dominant actors and structural characteristics, and Twitter to investigate the evolution of discourses and real-time reactions to various discrete events and processes. For the collection, analysis, and visualisation of networks, actors, and debates on AI ethics, the project will deploy the following tools: NodeXL visual representations and analytics for Twitter [Hansen, Schneiderman, Smith, 2010]; Netvizz, for data collection and extraction, an application tool allowing the export of data in file formats from different sections of the Facebook social networking service [Rieder, 2013], and Gephi open-source network graph and analysis tool [Cherven, 2015] for analysis and visualisation. These digital methods efficiently support the research objectives of the study with no need for engaging more complicated statistical and analytical tools which often require researchers with a rare social data science skillset (social science disciplinary background with the ability to use Python, R, UCINet).

Conclusion

The selection of a universal ethical standard on AI is the Holy Grail in this research area. Offering a holistic usable answer to this problem will benefit several disciplines, and stakeholders in global interdisciplinary academic and transnational practitioners fora. This research design proposed here has its high gains and high risks. First, the intercontinental scope of the study. It involves the collection, integration and expansion of material for a new framework for AI ethics. Although it might not produce the ultimate answer of the absolute standard, it will generate new data, findings, and advanced detailed recommendations taking AI ethics to new lines of inquiry. Second, data collection. It is a recognised possibility that we will not have access to specific individuals to interview in the actors already identified. This risk could be addressed by relying on researchers' networks to locate interview participants. Third, it involves

DOI: 10.14515/monitoring.2021.1.1911

fieldwork around the world, which contains a certain level of risk for researchers; however, the countries specified are not presently volatile to political developments and social unrest. Interviews can also be conducted online to mitigate travel risk in relation to the COVID-19 pandemic developments. Last, the fieldwork's operational success relies on solid research management to co-ordinate the five work packages reflecting the POs, so that they feed seamlessly into each other, enhanced by frequent reviews, peer assessment, and reports, as well as mentoring, collegiality, and sensitivity in managing the research teams involved.

References (Список литературы)

Acemoglu D., Restrepo R. (2018) The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment. *American Economic Review*. Vol. 108. No. 6. P. 1488—1542. https://doi.org/10.1257/aer.20160696.

Bostrom N. (2014) Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press.

Cherven K. (2015) Mastering Gephi Network Visualization. Birmingham (UK): Pack Publishing Limited.

Chouldechova A. (2017) Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*. Vol. 5. No. 2. P. 153—163. https://doi.org/10.1089/big.2016.0047.

Codagnone C., Karatzogianni A., Matthews J. (2018) Platform Economics: Rhetoric and Reality in the "Sharing Economy". Bingley: Emerald Publishing Limited.

Floridi L. (ed.) (2015) The Online Manifesto: Being Human in a Hyperconnected Era. Cham: Springer.

Floridi L. (2018) Soft Ethics and the Governance of the Digital. *Philosophy & Technology*. No. 31. P. 1—8. https://doi.org/10.1007/s13347-018-0303-9.

Garfinkel S., Matthews J., Shapiro S. S., Smith J. M. (2017) Toward Algorithmic Transparency and Accountability. *Communications of the ACM*. Vol. 60. No. 9. P. 5. https://doi.org/10.1145/3125780.

Hansen D., Schneiderman B., Smith M. A. (2010) Analyzing Social Media Networks with NodeXL: Insights from a Connected World. San Francisco (US): Elsevier Science & Technology.

Harari Y. N. (2016) Homo Deus: A Brief History of Tomorrow. London: Vintage Publishing.

Hintz A., Dencik L., Wahl-Jorgensen K. (2018) Digital Citizenship in a Datafied Society. Oxford: Polity Press.

Jobin A., Ienca M., Vayena E. (2019) The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence*. No. 1. P. 389—399. https://doi.org/10.1038/s42256-019-0088-2.

DOI: 10.14515/monitoring.2021.1.1911

Kaplan J. (2015) Humans Need Not Apply: A Guide to Wealth and Work in the Age of Artificial Intelligence. New Haven: Yale University Press.

O'Neil C. (2016) Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. London: Penguin Books Ltd.

Redden J., Dencik L., Warne H. (2020) Datafied Child Welfare Services: Unpacking Politics, Economics and Power. *Policy Studies*. Vol. 41. No. 5. P. 507—526. https://doi.org/10.1080/01442872.2020.1724928.

Rieder B. (2013) Studying Facebook via Data Extraction: The Netvizz Application. In: *Proceedings of the 5th Annual ACM Web Science Conference (WebSci'13)*. May. P. 346—355. https://doi.org/10.1145/2464464.2464475 (accessed: 27.02.2021).

Sanchez-Monedero J., Dencik L., Edwards L. (2020) What Does it Mean to 'Solve' the Problem of Discrimination in Hiring? Social, Technical and Legal Perspectives from the UK on Automated Hiring Systems. Paper presented at Conference on Fairness, Accountability, and Transparency (FAT*'20). Barcelona, Spain. 27—30 January. URL: https://arxiv.org/abs/1910.06144 (accessed: 27.02.2021).

Scott J. (2000) Social Network Analysis: A Handbook (2nd ed.). London: Sage.

Sumpter D. (2018) Outnumbered: From Facebook and Google to Fake News and Filter-bubbles — The Algorithms that Control our Lives. London: Bloomsbury Publishing PLC.

Wasserman S., Faust K. (1994) Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press.

van Dijck J. (2014) Datafication, Dataism and Dataveillance: Big Data between Scientific Paradigm and Ideology. *Surveillance & Society*. Vol. 12. No. 2. P. 197—208. https://doi.org/10.24908/ss.v12i2.4776.