RECERCA, REVISTA DE PENSAMENT I ANÀLISI, NUM. 27(2). 2022 e-ISSN: 2254-4135 – pp. 1-26
DOI: http://dx.doi.org/10.6035/recerca.6147

The Banality of (Automated) Evil: Critical Reflections on the Concept of Forbidden Knowledge in Machine Learning Research

La banalidad del mal (automatizado): reflexiones críticas sobre el conocimiento peligroso en la investigación del aprendizaje automático

ROSA M. SENENT[1] (Dublin City University) and DIEGO BUESO (Universitat de València)

Article received: 15, September, 2021
Revision request: 12, January, 2022
Article accepted: 25, April, 2022

Senent, Rosa M. and Bueso, Diego (2022). The Banality of (Automated) Evil: Critical Reflections on the Concept of Forbidden Knowledge in Machine Learning Research. ***Recerca. Revista de Pensament i Anàlisi*****, 27(2), pp. 1-26.**
doi: http://dx.doi.org/10.6035/recerca.6147

Abstract

The development of computer science has raised ethical concerns regarding the potential negative impact of machine learning tools on people and society. In this article, we provide three examples of automated evil: deepfake technology (ab)used by anonymous men to make digitally manipulated pornography to harm women; pattern recognition designed to try to uncover sexual orientation; and deep learning and extensive datasets used by private companies to influence democratic elections. We contend that the concept of 'forbidden knowledge' can help to inform a coherent ethical framework in the context of data and computer science research and contribute to tackle automated evil. We conclude that restricting generalised access to extensive data and limiting access to ready-to-use codes would mitigate potential harm caused by machine learning tools. In addition, we advocate that the notions of intersectionality and interdisciplinarity be systematically incorporated in data and computer science research.

Key Words: AI ethics, machine learning, artificial intelligence, forbidden knowledge, gender.

[1] Corresponding author: rosa.senentjulian2@mail.dcu.ie

Resumen

El desarrollo de las ciencias computacionales ha suscitado preocupaciones éticas relacionadas con las consecuencias negativas de algunas herramientas de aprendizaje automático sobre las personas y la sociedad. En este artículo, ofrecemos tres ejemplos de mal automatizado: tecnología *deepfake* (ab)usada por hombres anónimos para hacer pornografía y dañar a mujeres; un modelo de reconocimiento de patrones diseñado para intentar descubrir la orientación sexual; y aprendizaje profundo y datos usados por compañías privadas para influir en elecciones democráticas. Defendemos que el concepto de *conocimiento peligroso* puede formar parte de un marco ético coherente en las ciencias computacionales y ayudar a reducir el mal automatizado. Concluimos que restringir el acceso generalizado a extensas bases de datos y limitar el acceso a códigos disponibles para su uso podría mitigar los daños causados por algunas herramientas de aprendizaje automático. Además, las nociones de interseccionalidad e interdisciplinariedad deberían incorporarse sistemáticamente en la investigación de ciencia de datos y computacional.

Palabras clave: ética, aprendizaje automático, inteligencia artificial, conocimiento prohibido, género.

INTRODUCTION

Science is not produced in a social vacuum. Rather, scientists conduct scientific research in a specific social, cultural, historical, and economic context which influences the type of topical and methodological dimensions of scientific production. For centuries, scientists have presented science as something objective and neutral. However, they are not immune to social biases and prejudices. Scientists are social agents with the capacity to deploy pseudoscientific arguments to defend the biological inferiority of women, black people, and other socially subordinated groups.[2] This alleged objectivity of science gave legitimacy to flawed claims made by the scientific community —almost exclusively made up of white men— that had harmful consequences for marginalised populations. Scholars in the fields of sociology and philosophy of science have long debunked the myth of scientific objectivity. There has been a growing acknowledgement that sciences, both the social sciences[3] and the

[2] At the turn of the century, a researcher claimed that the brain of the average "grown-up Negro partakes, as regards his intellectual faculties, of the nature of the child, the female, and the senile White" (cited in Kimmel, 2000: 30). As Kimmel himself notes, one can only speculate where this leaves older black women.

[3] For example, in the past, anthropologists concentrated on the male sector of the populations they studied, relegating women to the periphery of anthropological research (Milton, 1979; Reiter, 2012). In psycholo-

natural sciences,[4] have emerged in a patriarchal context and, as such, they have been constructed upon androcentric principles (Harding, 1996; Kimmel, 2000). It is now widely accepted that socially and historically situated individuals produce knowledge and that the production of knowledge is not an abstract reality, but a process rooted in social contexts.

The origins of computer science can be traced back to the middle of the twentieth century. Without the possibility of physical implementation, the theoretical basis of this newly born science was just another mathematical application. The automation era began only when the first computers were capable of solving basic logical problems and performing iterative functions. By the 2010s, the science of machine learning (ML), also known as artificial intelligence (AI), was advancing rapidly in terms of theoretical insights but was used primarily for practical applications. This can be considered a change of scientific paradigm in the Kuhnian sense (1971). The discovery of a new powerful knowledge base within a particular scientific field does not change the social character of this science. Like other scientists, researchers in the computer science community are not immune to social biases and prejudices.

'Artificial intelligence' is the term most frequently used in the media and academic publications. We consider that this term is problematic for a number of reasons. First, the ability to learn does not necessarily equate to intelligence. Second, this term, which suggests that unsupervised autonomy is possible in machines, could raise false expectations in the general public in relation to the tasks machines can actually perform. Third, although this notion can be misleading, it can be exploited by Big Tech companies to feed the science-fiction idea according to which they are selling truly intelligent machines to potential customers. Therefore, we use the term 'machine learning' which more accurately represents that of which it speaks, namely: the capability of a machine to learn and perform a specific task. Machine learning refers to what computer systems do when they are able to automate and generalise a

gy, androcentric biases include the underrepresentation of women as researchers and as research participants, and researchers' practices in comparing women and men and describing their research findings (Hegarty & Buechel, 2006; Eagly & Riger, 2014).

[4] Androcentric biases are also present in biomedical research, where experimental results obtained from research using only male participants have been extrapolated to females, hence compromising our understanding of female biology and adversely affecting women's health (Beery & Zucker, 2011; Lee, 2018). Eighteenth-century prejudices about the supposed inferiority of the female brain, which were used to justify the exclusion of women from the public sphere, are still prevalent in contemporary brain research, despite neuroscientists reporting no significant differences between the brains of women and men (Malane, 2005; Jordan-Young, 2011; Rippon, 2019).

task without following explicit programming scripts, by inferring models from data.

Machine learning models need to be fed data to be accurate in their predictions. The availability of extensive data, coupled with current computational power, are two of the most relevant factors at the heart of the machine learning revolution. In the last few decades, machine learning has progressed significantly thanks to the systematic collection of large amounts of personal data carried out by "privacy-invasive" interactive technologies, such as social media platforms and mobile phone apps (Hagendorff, 2020b: 110). Importantly, machine learning models, which "seem to be backed up by data and 'science'", are often not scrutinised because they are assumed to be objective (Gebru, 2019: 5).

The idea that computer science is neutral because it relies on 'artificially intelligent' algorithms that supposedly are blind to the social context in which computer scientists create them is an extension of "the myth of scientific objectivity" (Gebru, 2019: 5). The supposed neutrality of algorithms can be considered dangerous in itself. Far from "introducing objectiveness", machine learning algorithms can, in fact, "perpetrate discrimination" (Allen & Masters, 2020: 588). Indeed, "people worry that computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the world" (Domingos, cited in Tajalli, 2021: 449). In the current globalised context, the machine learning revolution has fomented an unprecedentedly close relationship between technology and society, a new paradigm under which the application of technologically advanced tools can have deeply disturbing social consequences.

1. THE BANALITY OF (AUTOMATED) EVIL

The concept of the "banality of evil" developed by political philosopher Hannah Arendt (1999) can shed light on this complex panorama, in which not only (supposedly neutral) algorithms can perpetuate discrimination, but socially irresponsible individuals can use algorithms for immoral purposes. In analysing the case of Adolf Eichmann, a German high official who took part in the Nazi massacre and who allegedly was just following orders, Arendt argued that evil may be caused by an absence of critical thought rather than by an individual predisposition to evil. She used the phrase "banality of evil" to refer to cases in which people refuse to engage in critical thinking and blindly

follow their own predetermined values or obey external orders, which results in their committing monstrous acts (Tajalli, 2021: 450). She posited that evil was a consequence of the refusal to think on the part of many individuals who, like Eichmann, uncritically obey orders. According to Arendt, the ability to apply critical thinking to situations involving moral conflict is a crucial requirement to avoid the blind perpetration of evil acts.

This consideration can also be applied to the context of "rule-abiding" and "conformist" algorithms which blindly follow orders given by their potentially biased programmers or arrive by themselves at morally problematic decisions due to their lack of human critical thinking (Tajalli, 2021: 451). Algorithms 'obey orders' given by a code and neither think about nor question the social consequences and the ethical implications of their doing so. Machine learning tools are "ethically blind" because, despite their ability to process large volumes of data, they lack the type of intelligence required to be moral agents (Tajalli, 2021: 451). This also explains why "ethical decision-making cannot be reduced and be based on a set of codes or pre-defined algorithms" (Tajalli, 2021: 448). The deployment of technology that has learned social biases can cause harm to vulnerable populations (Bender, Gebru, McMillan-Major & Schmitchell, 2021: 615). Also, individuals with morally questionable purposes can employ machine learning tools to reinforce biases against marginalised people. Therefore, the banality of automated evil can have dangerous outcomes that need to be critically addressed by ethicists and philosophers, and prevented by data and computer scientists.

2. SOCIALLY IRRESPONSIBLE USES OF MACHINE LEARNING

The computer science community has arrived at a consensus to apply guidelines of good practice that prevent practical problems in some algorithms. This has permitted successful positive applications of machine learning in many research fields not directly connected to computer science. To give just a few examples: machine learning was applied to find efficient ways of reducing our dependence on fossil fuels to tackle the consequences of climate change (Rolnick et al., 2019); deep learning can provide accurate long-term forecasts of climate anomalies that cause droughts and floods (Ham, Kim & Luo, 2019); and, recently, an algorithm was capable of predicting the risk of heart failure better than human experts (Alaa et al., 2019). Also, in computer sciences, some researchers focus on preventing user influence on models. Such

is the case of fair learning, which tries to represent all types of data equally in models (Feuerriegel, Dolata & Schwabe, 2020).

A number of issues linked to the negative consequences of various machine learning applications have been raised by experts in the emerging field of "AI ethics" (Jobin, Ienca & Vayena, 2019). It has been found that social notions of race and gender affect both the design and usage of machine learning-based systems, perpetuating societal prejudices held by the overwhelmingly male[5] and white[6] individuals behind the machines (Buolamwini & Gebru, 2018). For example, the Internet-based training data for deep learning models (for example those employed for computer vision applications) often have a number of problematic characteristics which "overrepresent hegemonic viewpoints" and result in "models that encode stereotypical and derogatory associations along gender, race, ethnicity, and disability status" which "amplify biases and harm" for marginalised populations (Bender, Gebru, McMillan-Major & Schmitchell, 2021: 615). In the last few years, books such as *Weapons of Math Destruction: how big data increases inequality and threatens democracy* (2017) by Cathy O'Neil and *Automated Inequality: how high-tech tools profile, police and punish the poor* (2019) by Virginia Eubanks have exposed how people in lower socio-economic classes in the United States are "subjected to more automated decision-making tools than those who are in the upper class", with algorithms "most often used on people towards whom they exhibit the most bias" (Gebru, 2019: 1). Indeed, some technologically advanced tools can harm certain groups of people. However, the rapid implementation of machine learning tools in different scenarios has not been thoroughly investigated (Gebru, 2019: 5).

In the next section, we provide three examples of dangerous applications of machine learning tools which we consider different cases of automated evil. The first example illustrates how a huge number of anonymous men can (ab)use a specific type of technology to harm women; this constitutes a type of technology-based misogynistic violence. The second example deals with the misuse of technology on the part of researchers who carry out an ethically problematic type of research; the very design of the study is flawed, and the authors do not consider the dangerous implications that their study could

[5] From 2015 to 2018, 22% of professionals in the world of AI and ML were female (World Economic Forum, 2018). Globally, 17% of computer science graduates in 2019 were female, and 26% of professionals who had data and AI-related jobs in 2020 were female (Young, Wajcman & Sprejer, 2021).

[6] In the United States, 18% of individuals conducting doctoral research in computer sciences are female; 45% are white, while 22.4% are Asian, 3.2% Hispanic, and 2.4% African American (Stanford University's Institute for Human-Centered Artificial Intelligence, 2021).

have on homosexual people due to homophobia. The third example shows how private companies can harm society as a whole by deploying machine learning to influence democratic elections; this is a type of manipulation on a greater scale. After analysing these three examples, we contend that the application of the concept of forbidden knowledge could help to tackle automated evil in the context of data and computer science research.

2.1 Deepfake Technology: A Weapon of War Against Women

The possibility of open access[7] leaves the door open for some machine learning applications to be easily misused, which can lead to extremely harmful consequences (Hagendorff, 2020a). Such is the case of deepfakes, which depict people saying and doing things that never happened. Deepfakes use facial mapping technology and machine learning tools to replace the face of a person on a video with the face of another person. The resulting fake videos look genuine and spread quickly on social media, where they can be watched by thousands of people. Since the software necessary for crafting realistic deepfakes is available as open source,[8] growing numbers of digitally educated people, as well as those with few technical skills, can use this technology to harm others.

The consequences of technologically advanced tools such as deepfakes cannot be uncoupled from the socio-historical conditions in which they emerge. Structural inequality increases the likelihood of vulnerable groups being harmed by certain technologies (Gebru, 2019). Historically, women have been subordinated in Western patriarchal societies (Walby, 1992; Kimmel, 2000). Nowadays, the pornography industry is a multi-billion dollar business that benefits from the sexual exploitation of women while legitimising and normalising misogynistic violence (Dines, 2010). Addressing the main use of deepfake technology must take into consideration the current social context, in which women are devalued and objectified at an unprecedented level due to globalisation.

[7] Meaning scientists can openly share the details of their machine learning models.

[8] Deepfake creation tools rely on open-source code repositories, as do the great majority of machine learning projects. In 2017, "the faceswap source code from the anonymous /r/ Deepfakes creator was donated to the open-source community and uploaded on Github" (Ajder, Patrini, Cavalli & Cullen, 2019: 4). Since then, thousands of anonymous 'Eichmanns' have contributed to the banality of automated evil through the development of deepfakes.

The origins of deepfakes are closely entwined with pornography. The first deepfakes went viral in 2017 when digitally manipulated pornography was uploaded to a discussion website (Maddocks, 2020). Today, deepfakes are widely used to produce digitally manipulated pornographic videos. The connection between deepfakes in general and pornography is notable, as sexually abusive deepfakes can also be used for political blackmail (Kikerpill, 2020). The use of deepfake technology both for political purposes and for pornographic videos has been found to operate in a similar way to silence critical speech (Maddocks, 2020). For example, when a female journalist in India started to uncover government corruption, her face was grafted onto a pornographic video; the video went viral, and she was violently harassed and subjected to rape threats to the point that she had to go offline for several months (Hao, 2019). The use of deepfake technology disproportionately targets women; it is overwhelmingly (ab)used by men for the purpose of harming women, as in the case of revenge porn.[9] A recent report on the state and impact of deepfakes shows that 96% are pornographic and that 100% of victims are female (Ajder, Patrini, Cavalli & Cullen, 2019: 5). The consequences of deepfakes deployed as a misogynistic weapon are still under-researched (Maddocks, 2020).

The use of pornographic deepfakes constitutes a type of "image-based sexual abuse"[10] (Henry et al., 2020), and, therefore, a form of "technology-facilitated gender-based violence" (Dunn, 2020). The consequences for women can be "life-ending, often devastating, relentless and isolating" (Rackley et al., 2021). One survivor defined her extreme emotional distress as "torture for the soul" (McGlynn et al., 2021: 557). The gendered nature of this technology-facilitated abuse can be better understood as part of a "continuum of sexual violence" (Kelly, 1987; Russell 1990), as the victims-survivors experience it as part of a broader context in which misogynistic violence, harassment, and abuse are rampant (Henry et al., 2020). This machine learning tool poses a major threat to human rights, and the very nature of this technology entails

[9] Revenge porn is understood as "the non-consensual distribution of private, sexual images by a malicious ex-partner" (McGlynn & Rackley, 2017: 26).

[10] Image-based sexual abuse is "the non-consensual taking, sharing or threats to share nude or sexual images of a person", which includes digitally manipulated images as in the case of pornographic deepfakes (Powell, Scott, Flynn & Henry, 2020: 2). This constitutes "a violation of [women's] human rights to dignity, sexual autonomy, and freedom of expression" (Powell, Scott, Flynn & Henry, 2020: 2).

many difficulties for the legal and psychological protection of its victims[11] (Rackley et al., 2021).

Many researchers have warned against the negative uses to which deepfake technology can be put (Westerlund, 2019; Kikerpill, 2020). The potential positive uses of deepfakes, highlighted nonetheless by some authors (Westerlund, 2019: 41) do not seem, in the slightest, to counterbalance the deeply negative consequences of their weaponisation. The negative applications outweigh the potential positive applications. Due to the generalised unethical application and the dangerous consequences of this technology, this constitutes a case of automated evil. Harm has already been done, but it is imperative to search for the means to prevent future harm. For the time being, both young and adult users should be taught how to detect deepfake technology[12] in order to minimise its negative impact. It is increasingly important to improve people's media literacy and critical thinking.[13] Urgent measures must be put in place against deepfakes, both legal measures and machine learning anti-deepfake tools (Westerlund, 2019). Governments and institutions must recognise deepfake-related violence as another type of gender-based violence, and institutional support must be provided to help the victims overcome the harm perpetrated against them. Above all, measures should be taken to hold the perpetrators accountable.

2.2 Pattern Recognition: A Violation of the Human Right to Privacy

Pattern recognition is a group of algorithms that extract information and knowledge from patterns and their representations. These algorithms try to classify data based on statistical information. They have been exploited with machine learning tools, such as neural networks, to classify facial photographs, what has become known as face recognition (Le, 2011). Some applications of this technology raise major ethical concerns. For example, the Chinese government uses this technology to monitor the movements of citizens and assess

[11] Legislation needs to properly address this issue. Currently, litigation is difficult and costly for victims (Rackley et al., 2021).

[12] A number of subtle indicators can aid people to detect deepfakes. These range from "unnatural eye direction" to "strange behaviour of an individual doing something implausible" (Westerlund, 2019: 45).

[13] Leaving such an important task in the hands of private corporations runs the risk of promoting an uncritical indoctrination on the benefits of machine learning development based on economic interests and justified by neoliberal ideology (Benkler, 2019). Education in computer sciences for the general public should primarily be to raise awareness and critically address the social consequences and ethical implications derived from the misuses of machine learning technology.

their behaviour, and then give them 'social credits' which can influence the outcome of a job interview and their access to public transport (Wong & Dobson, 2019). Face recognition can lead to cases of direct discrimination, in which individuals are "treated less favourably because of a protected characteristic" (Allen & Masters, 2020: 593). Also, some studies have shown "that pattern recognition technology can unintentionally lead to the replication of human biases in various subtle ways" (Allen & Masters, 2020: 588).

Here, we focus on a study that claims face recognition can detect sexual orientation (Wang & Kosinski, 2018). Using pattern recognition in facial photographs, the authors uncover physiological patterns linked to sexual orientation. The uncovered patterns refer to the most likely face of homosexual and heterosexual males and females within a specific group of the white population in the United States. However, these authors reach a number of highly problematic pseudo-scientific conclusions.

The authors blindly subscribe to "prenatal hormone theory",[14] insisting that PHT explains sexual orientation, "gender-atypical facial morphology, expression, and grooming styles" of homosexual men and women (Wang & Kosinski, 2018: 247). In their opinion, their results show that homosexual men have "feminine" and lesbian women "masculine facial features". However, the neural network they trained uncovered stereotypical features linked to the *social* category of gender and makes dramatic generalisations based on mean facial differences found in the four pre-selected categories on which they focus: heterosexual males, heterosexual females, homosexual males, homosexual females. For example, they claim to have found that heterosexual males have more facial hair than homosexual males, that homosexual males have longer noses, that heterosexual females use more makeup than homosexual females, and that homosexual females "smile less" than heterosexual females (Wang & Kosinski, 2018: 252). They state that heterosexual men and lesbian women tend to wear "baseball caps"[15] (Wang & Kosinski, 2018, 252) and make highly questionable claims such as that testosterone influences "dominance"[16] (Wang & Kosinski, 2018: 246).

[14] The authors contend that, "according to the PHT, same gender sexual orientation stems from the underexposure of male fetuses or overexposure of female fetuses to androgens that are responsible for sexual differentiation" (Wang & Kosinski, 2018: 247). For information on how hormones have often been used to justify the status quo, see Rippon (2019).

[15] Wearing baseball caps is not linked to prenatal hormone overexposure or underexposure, but to the specific cultural context of the United States, where baseball is a popular game.

[16] This sociobiological myth has long been debunked by anthropological, sociological, and feminist research on the causes of men's social domination (e.g., Bourdieu, 2000; Kimmel, 2000).

Their conclusions are based on a deep neural network which accurately describes gender stereotypes in a specific portion of the population (Agüera y Arcas, Todorov & Mitchell, 2018). A deeper conclusion is that correlation (the supposed relation between someone's face and their sexuality) does not imply causation (since people's environment and habits affect their faces). A *correlation* between two variables is not enough to assert a *causal* relation between them. These authors' experiments are designed in such a way that they reinforce gender-based stereotypes. This is a paradigmatic case of "essentialist research" as defined by Kimmel (Kimmel, 2000: 45), which links homosexuality with "gender inversion", as though women were the reference point against which homosexual and heterosexual men should be measured.

This study has a number of highly problematic implications. First, bisexuals are blatantly ignored, probably because bisexuality poses a problem when trying to link and essentialise gender roles, biological sex, and sexual orientation. Considering only the dichotomic stereotypical features linked to heterosexuals and homosexuals undoubtedly makes it easier to generalise gender-based stereotypes and heterosexist prejudices. Thus, this study is a case of biphobia. Biphobia can have devastating consequences for the lives of bisexual people (Welzer-Lang, 2008; Robinson, 2019). Second, their data samples are exclusively based on the faces of white people in the United States (Wang & Kosinski, 2018: 255). Despite that, they insist that their results could be extrapolated to other populations (Wang & Kosinski, 2018: 254), hence demonstrating a lack of awareness of the impact that biased, limited data can have on the output of experiments and investigations. Third, they themselves recognise that exposing sexual orientation could have serious and even life-threatening implications for homosexual women and men (Wang & Kosinski, 2018: 247). However, worryingly enough, they suggest that "publicly available data and conventional machine-learning tools" could be employed "to build accurate sexual orientation classifiers" (Wang & Kosinski, 2018: 255). In an interview, Michal Kosinski, one of the authors, spoke about the application of pattern recognition to extract vulnerable social data: "This is not my fault. I did not build the bomb. I only showed that it exists" (Grassegger & Krogerus, 2017). By refusing to engage in critical thinking and moral responsibility, he reveals himself as a representative banal perpetrator of evil.

To conclude, this is an example of how socially irresponsible researchers carry out an investigation with deeply disturbing implications for marginalised groups using machine learning tools in an inappropriate and unethical manner. Trying to expose a protected characteristic such as a person's sexual

orientation is a task that should not be undertaken in the first place, particularly when that characteristic is currently used as a justification for violence and even murder in many countries around the world (Ramón Mendos et al., 2020). This phrenology-like study is a paradigmatic case of automated evil. Researchers must abide by human rights principles and not jeopardise the safety of vulnerable groups.

2.3 Misuse of Social Data and Deep Learning: A Threat to Democracies

Extensive datasets have prompted the development of a new range of social models to make predictions. These models can be used by governments and private companies to assess strategies to disseminate messages. In 2019, the Global Inventory of Organised Social Media Manipulation stated: "Social media, which was once heralded as a force for freedom and democracy, has come under increasing scrutiny for its role in amplifying disinformation, inciting violence, and lowering levels of trust in media and democratic institutions" (Bradshaw & Howard, 2019). Since 2014, the number of democratic countries in which social media and machine learning tools have been used in elections and referendums has grown significantly (Pastor-Galindo et al., 2020). The procedure in almost every case is similar, starting with the extraction of social statistical information from individuals to train a model. This model is then used to predict individuals' responses to certain messages (Youyou, Kosinski & Stillwell, 2015).[17] Using bots[18] as amplifiers, specific messages are spread in accordance with people's individual characteristics in order to improve the chance of success.

The most striking case was that of the private company Cambridge Analytica, which influenced the 2016 presidential elections in the United States. In 2015, Cambridge Analytica proved that 12 'likes' of a Facebook user were enough to predict personal characteristics such as depression, impulsivity, and life satisfaction (Youyou, Kosinski & Stillwell, 2015). Before the elections, Cambridge Analytica extracted personal information from the Facebook accounts of almost 60 million users in the United States, with the consent of Facebook. This extensive dataset was used to train a social model to spread messages of political groups with an alt-right ideology, linked to candidate Donald Trump. Similar cases happened in the Brexit referendum in the Unit-

[17] Among the authors of this study is the above-mentioned Michal Kosinski.

[18] A bot is a social media account ruled by an algorithm to interact with other users and spread messages.

ed Kingdom in 2016 (Howard & Kollanyi, 2016) and in the general elections in Spain in 2019 (Pastor-Galindo et al., 2020). This reveals how sensitive democratic societies are to the exposure of personal information to companies with evil intentions. Therefore, this type of automated evil constitutes a threat to the trust of citizens in democratic elections and hampers social cohesion.

3. FORBIDDEN KNOWLEDGE

3.1 Forbidden Knowledge in Sciences

Forbidden knowledge refers to knowledge considered to be "too sensitive, dangerous or taboo to be produced or shared" (Hagendorff, 2020). A better term would be 'dangerous knowledge'.[19] Science is an active process, socially and historically situated, and as such belongs in the domain of the social, the moral, and the political (Johnson, 1996: 197). Thus, the question of forbidden knowledge in science is a moral, social, and political issue (Johnson, 1996: 206). This concept has been applied to some lines of scientific research which may produce a degree of harm that outweighs its potential positive effects.

Forbidden knowledge is about deciding where science should go as well as where it should not go (Johnson, 1999: 453). Because the applications of certain types of knowledge have an impact on the world, a secularised notion of forbidden knowledge applied to science can be of help to determine which types of knowledge are worth pursuing and which are not. For this purpose, it is useful to characterise the aim of science as "knowledge for the good of humanity" (Johnson, 1996: 212). Hence, the question of forbidden knowledge is more a matter of forbidding certain lines of inquiry in cases in which a particular type of knowledge goes against the main aim of science: the good of humanity (Johnson, 1996: 213). Forbidding specific lines of inquiry that can have a negative impact on people and societies is, therefore, a way of fostering responsible science (Johnson, 1996: 213).

3.2 Forbidden knowledge in Machine Learning Research

Although the production of science is always guided by (and charged with) specific social values, this is usually "left out of the picture" (Johnson,

[19] Throughout this article, we refer to 'forbidden knowledge' because this is the most commonly used term.

1999: 452). Currently, scientific research has norms and practices that prohibit certain behaviours, such as fraud and plagiarism, and forbid experiments that mistreat animals and deceive individuals. These practices are methodological, rather than topical, but they illustrate how certain prohibitions are in place to foster responsible science (Johnson, 1996: 214). There are also "unspoken rules" shared by the scientific community as to what is considered forbidden knowledge; as one interviewee stated, "every microbiologist knows not to make a more virulent pathogen" (Kempner, Perlis & Merz, 2005: 854). We contend that the "social-moral-political argument" (Johnson, 1996: 206) put in favour of forbidden knowledge in other sciences should be applied to machine learning research. This concept should be part of the ethical framework needed in computer science (Hagendorff, 2020a).

Scientists must be willing to apply critical thinking and act in an ethically responsible way to foster ethical research, but the pressure to publish in order to advance their careers can be counterproductive.[20] One way of fostering precautionary measures to protect forbidden knowledge in machine learning research (and in other sensitive types of research) would be for academic institutions to not make the number of publications a matter of prestige and career advancement. The pressure to publish jeopardises the quality of scientific knowledge (Sarewitz, 2016). The lack of pressure would leave researchers free to carefully analyse the results of their research projects and experiments, assess whether the outcomes would, overall, be positive or negative for people and society, and make an ethically responsible decision regarding publication.

4. PROPOSALS TO TACKLE AUTOMATED EVIL: FORBIDDEN KNOWLEDGE, INTERSECTIONALITY & INTERDISCIPLINARITY

In this section, we make a number of proposals and suggestions which could help to mitigate the harm derived from the irresponsible use of machine learning tools.

[20] Also, the pressure to publish promotes epistemic inequality and gives rise to a number of predatory practices in line with neoliberal attitudes in the context of academia (for instance, academic journals that charge authors to publish). This often turns the dissemination of knowledge into a matter of social class and economic power, and provides a way to buy academic status and prestige.

4.1 Promotion of Forbidden Knowledge

The machine learning research community embraces the notion of open access. In many cases, this can be beneficial. However, as illustrated by our earlier example of deepfake technology, fully open access to machine learning resources precludes the possibility of taking "precautionary efforts to prevent the malicious use of machine learning applications" (Hagendorff, 2020a: 3).

Different levels of knowledge represent different potential hazard levels. Expertise in machine learning is dependent on knowledge of how algorithms work (mathematical knowledge, coding skills) and the ability to employ real data in practical applications. On the one hand, being able to build an algorithm does not enable users to acknowledge data bias. On the other hand, using the code of an algorithm at the user level developed without knowing the internal process may produce unreliable results. Limiting access to certain types of knowledge is crucial to prevent the misuse of machine learning.

As a point of reference, we want to highlight the case of nuclear physics in the twentieth century, where the notion of forbidden knowledge was employed to halt the development of nuclear weapons (Smith, 1978). Knowledge of nuclear fission can be considered "threatening to our material well-being insofar as it could lead to material devastation of the planet" (Johnson, 1996: 203). The concept of forbidden knowledge was applied in Iran in 2015, with the Joint Comprehensive Plan of Action (JCPOA) which limits and monitors the production of enriched uranium[21] in order to prevent Iran from developing nuclear weapons. Inspired by this approach, we present two key points which could be applied to lower the risk of unethical applications in the context of machine learning research.

Formally, two elements are needed to run a machine learning algorithm correctly: the processed data and the code.

Processed data. Practical knowledge is conditioned to the availability of well-processed data. Without sufficient data, machine learning models cannot be trained effectively. Access to extensive and social datasets should be limited in order to avoid a situation in which *any* model can be trained by *anybody*, leading to harmful consequences for people and communities. Certain datasets are already restricted by numerous governments, institutions, and companies, which require users to register and provide their personal information

[21] Enriched uranium is uranium with a high concentration of isotope 235, needed for the fission process of uranium.

before they are granted access to extensive databases. This keeps the traffic and availability of large social datasets under control. As illustrated by the example of Cambridge Analytica, access to extensive datasets can pose a threat to social cohesion and trust. A more generalised restriction of vulnerable data could promote both the legal accountability and the social responsibility of social agents using this technology.

Code. Theoretical knowledge of computer science is needed to develop new tools, but easy access to the code makes mastering the basics of machine learning unnecessary. The integrity of the code is a crucial aspect in this regard. A neural network is easy to code, and there are many repositories and libraries where it can be used easily. But a neural network needs expertise because it is significantly sensitive to changes in data, architecture, and initial parameters (Heaven, 2019). Preventing access to the full code, or just allowing access to a demo for identified users, are measures that can be taken in order to keep control over ready-to-use codes. This has been applied in the cases of GPT-3[22] and Face2Face, where the complete codes were not released (Hagendorff, 2020a). This would be particularly useful to prevent the current generalised and dangerous use of deepfake technology. We have to clarify that without the support of governments, companies, and researchers who have access to well-scripted codes and data, none of these limitations would be effective.

4.2 Intersectionality

Machine learning applications must not be a threat to human dignity. Although human rights law and principles are not yet ready to cover all the potential negative social impacts of machine learning applications, they can be of help in this regard (Latonero, 2018: 24). While "seeking harmonization between AI ethics codes (soft law) and legislation (hard law) are important next steps for the global community" (Jobin, Ienca & Vayena, 2019: 8), it is also important to educate computer scientists in the ethical implications and social consequences of certain lines of investigation. Crucially, the computer science community must respect and protect fundamental human rights. As illustrated by the example of the authors who tried to uncover a protected

[22] In June of 2020, Open AI decided to release GPT-3 along with a waiting list in order to limit its usage. However, in November 2021, GPT-3 became fully open access. OpenAI's API Now Available with No Waitlist. *OpenAI*. Retrieved from: https://openai.com/blog/api-no-waitlist/ [Consulted 10 April 2022].

characteristic (sexual orientation), researchers must be trained to deal with the ethical implications of their experiments in an appropriate way. Data and computer scientists should be given guidelines to guarantee human rights, and measures should be put in place to ensure that machine learning research does not entail risks for vulnerable populations. As part of the process of upholding human rights, data and computer science researchers would benefit greatly from applying the feminist principle of intersectionality to the design and application of their investigations.

The concept of intersectionality developed by feminist scholar Kimberlé Crenshaw (1991) in critical race theory is crucial for weighing the consequences of machine learning tools, their applications, and consequences (Gebru, 2019: 11). The notion of intersectionality emphasises the social interaction between race, sex/gender, and class in the oppression of certain groups of people. This concept is useful to understand the intimate interrelation between oppressive categories within society, which are embedded in institutions such as the school, the family, and the state. The life experience of a given individual is dependent on the number of vulnerable groups the individual belongs to. For example, a migrant, working-class black lesbian or bisexual woman from Brazil living in the United States does not have the same social opportunities (as regards education, employment, and physical safety in the streets) as does an upper-class white heterosexual man from England living in the capital of his own country. In the context of computer science research, thanks to intersectionality we can understand why darker females tend to be underrepresented in machine learning datasets in relation to white males (Buolamwini & Gebru, 2018), and why most datasets (including datasets with medical applications) originate in North America and Europe (Khan et al., 2021) and not in Latin America or Africa. Applying the notion of intersectionality can help to protect fundamental human rights. Thus, intersectionality should be applied transversely in *all* areas of research.

4.3 Interdisciplinarity

The lack of interdisciplinarity in research is highly problematic. Interdisciplinarity is a significant cognitive challenge for researchers who explore issues beyond their field of expertise. This challenge is related to the epistemic divisions between the disciplines, the intellectual access to the literature of other fields, and the ability to perform effective peer review of work written from the perspective of other disciplines (Baum, 2020). Many studies have

recently pointed to interdisciplinarity as an environmental improvement for machine learning practices (Baum, 2020; Kusters et al., 2020; Viseu, 2015). The cognitive gap between research disciplines is immense and interdisciplinarity is difficult (Viseu, 2015). Therefore, there is an important role for intermediate-term interdisciplinarity systems that could make major contributions to address social problems (Baum, 2020). Here, following Nissani (1997), we highlight three illuminating points in this regard: first, scholars working within a particular discipline ("disciplinarians") often commit errors which can be best detected by people familiar with the different perspectives offered by two or more disciplines ("interdisciplinarians"), thus ensuring a more objective approach; second, interdisciplinarians may help breach communication gaps within academia, hence mobilising intellectual resources in order to achieve greater social rationality and justice; third, by bridging fragmented disciplines, interdisciplinarians can play a role in the defence of academic freedom (Nissani, 1997: 201). Promoting synergic work between computer sciences and social sciences has a direct impact on machine learning development and its influence on society. This paper aspires to make a contribution to this type of synergic work.

Science is not only social action, but a social institution with the power to either condone or condemn certain types of knowledge. The knowledge produced by scientific inquiry is inevitably shaped by the system in which it is produced (Johnson, 1996: 207). When institutions and governments offer funding in certain fields of research, they are directly influencing what knowledge is created (Johnson, 1999: 452). Private companies have economic interests in the development of certain types of technologies, and they can invest private funding for their own benefit (Benkler, 2019). Thus, public institutions must promote incentives to support the creation of interdisciplinary groups which would be encouraged to forge collaborations that ensure ethical machine learning research. For example, the European Research Council (ERC) Synergy Grant directly promotes interdisciplinarity by supporting and funding research projects in which experts from different disciplines participate. Ethics committees composed of different experts (within the industry, in university departments, etc.) and collaborations between experts from different disciplines could also help to ensure that "automated decision tools are created by people from diverse backgrounds" who understand "the historical and political factors that disadvantage certain groups who are subjected to these tools" (Gebru, 2019: 1). These interdisciplinary collaborations should be

positively valued, as they would provide us with the opportunity to create socially responsible data and computer science.

5. CONCLUSION

In the current global context, where neoliberal capitalism has become the hegemonic economic and social model due to globalisation, money-driven corporations have the power to promote a generalised and uncritical approach that sets no limits whatsoever to data and computer science research. Undoubtedly, the economic interests linked to the development of these technologies can hamper ethical behaviour in this context. Given the potential of certain machine learning applications to negatively impact individuals and society, an uncritical attitude within this field can only bring about immense harm to even more people and communities. Since machine learning technology can have dangerous consequences, computer science research entails socio-political issues that must be acknowledged, analysed, and addressed. Therefore, the data and computer science community is in urgent need of an ethical framework that can help its members to deal with what we have called the banality of automated evil.

In this article we have commented on a number of cases that have harmed (and can continue to harm) people and communities at different social levels. The first example analyses deepfake technology, which is primarily used as a weapon of war against women through digitally manipulated pornography made by men. The second example exposes some of the dangerous implications of a study that claims that facial recognition can detect sexual orientation. The third case illustrates how social data and machine learning can be used to influence elections in democratic countries, hence threatening social cohesion and trust. Major measures should be taken to reduce the potential misuse of machine learning tools which come under the definition of forbidden knowledge, since the consequences of this misuse are too dangerous for people and society. We contend that the application of the concept of forbidden knowledge to certain lines of research, which are currently pursued uncritically in the field of data and computer science, is not only necessary but urgent in the name of responsible science. This is also a matter of social justice. We argue that limiting generalised access to extensive data, as well as limiting access to ready-to-use codes, would mitigate the effects of automated evil. In addition, in order to foster ethics and human rights, we advocate that

intersectionality and interdisciplinarity be systematically incorporated into data and computer science research.

BIBLIOGRAPHY

Agüera y Arcas, Blaise, Todorov, Alexander & Mitchell, Margaret (2018). Do algorithms reveal sexual orientation or just expose our stereotypes? *Medium*. Retrieved from: https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477 [Consulted 20 August 2021].

Ajder, Henry, Patrini, Giorgio, Cavalli, Francesco & Cullen, Laurence (2019). *The state of deepfakes: landscape, threats, and impact*. Amsterdam: Deeptrace.

Alaa, Ahmed, Bolton, Thomas, di Angelantonio, Emanuele, Rudd, James H. F. & van der Schaar, Mihaela (2019). Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLOS ONE*, 14(5), 1-17.

Allen, Robin & Masters, Dee (2020). Artificial Intelligence: the right to protection from discrimination caused by algorithms, machine learning and automated decision-making. ERA Forum 20. *Springer*, 585-598.

Arendt, Hannah (1999). *Eichmann en Jerusalén: un estudio sobre la banalidad del mal*. Barcelona: Lumen.

Baum, Seth (2020). Artificial Interdisciplinarity: Artificial Intelligence for Research on Complex Societal Problems, Philosophy & Technology. Retrieved from: https://ssrn.com/abstract=3651313 [Consulted 13 July 2021].

Beery, Annaliese & Irving Zucker (2011). Sex Bias in Neuroscience and Biomedical Research. *Neuroscience & Biobehavioral Reviews*, 35(3), 565-572.

Bender, Emily, Gebru, Timnit, McMillan-Major, Angelina & Schmitchell, Schmargaret (2021). On the Dangers of Stochastic

Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (610-623).

Benkler, Yochai (2019). Don't let industry write the rules for AI. *Nature*, 569(7754), 161.

Bourdieu, Pierre (2000). *La dominación masculina.* Barcelona: Anagrama.

Bradshaw, Samantha & Howard, Philip (2019). *The global disinformation order: 2019 global inventory of organised social media manipulation*. Oxford: Project on Computational Propaganda.

Buolamwini, Joy & Gebru, Timnit (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on fairness, accountability and transparency*. PMLR, 77-91.

Crenshaw, Kimberle (1990). Mapping the margins: Intersectionality, identity politics, and violence against women of colour. *Stanford Law Review*, 43(6), 1241-1299.

Dines, Gail (2010). *Pornland: How porn has hijacked our sexuality*. Boston: Beacon Press.

Dunn, Suzie (2020). Technology-Facilitated Gender-Based Violence: An Overview. *Centre for International Governance Innovation: Supporting a Safer Internet*, 1.

Eagly, Alice & Riger, Stephanie (2014). Feminism and psychology: Critiques of methods and epistemology. *American Psychologist*, 69(7), 685-702.

Feuerriegel, Stefan, Dolata, Mateusz & Schwabe, Gerhard (2020). Fair AI. *Business & Information Systems Engineering*, 62(4), 379-384.

Gebru, Timnit (2019). Race & Gender. In *Oxford Handbooks of AI Ethics.* Oxford: Oxford Handbooks.

Grassegger, Hannes & Krogerus, Mikael (2017). The Data That Turned the World Upside Down. *Vice*. Retrieved from: https://www.vice.com/en/article/mg9vvn/how-our-likes-helped-trump-win [Consulted 21 August 2021].

Hagendorff, Thilo (2020a). Forbidden knowledge in machine learning reflections on the limits of research and publication. *AI & SOCIETY*, 1-15.

Hagendorff, Thilo (2020b). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30(1), 99-120.

Ham, Yoo-Geun, Kim, Jeong-Hwan & Luo, Jing-Jia (2019). Deep learning for multi-year ENSO forecasts. *Nature*, 573(7775), 568-572.

Hao, Karen (2019). An AI app that "undressed" women shows how deepfakes harm the most vulnerable. *MIT technology review*. Retrieved from: technologyreview.com/2019/06/28/134352/an-ai-app-that-undressed-women-shows-how-deepfakes-harm-the-most-vulnerable/ [Consulted 23 August 2021].

Harding, Sandra (1996). *Ciencia y feminismo*. Madrid: Morata.

Heaven, Douglas (2019). Why deep-learning AIs are so easy to fool. *Nature*, 574(7777), 163-166.

Hegarty, Peter & Buechel, Carmen (2006). Androcentric Reporting of Gender Differences in APA Journals: 1965-2004. *Review of General Psychology*, 10(4), 377-389.

Henry, Nicola, McGlynn, Clare, Flynn, Asher, Johnson, Kelly, Powell, Anastasia & Scott, Adrian J. (2020). *Image-based Sexual Abuse: A Study on the Causes and Consequences of Non-consensual Nude or Sexual Imagery*. London: Routledge.

Howard, Philip & Kollanyi, Bence (2016). Bots,#strongerin, and #brexit: Computational propaganda during the UK-EU referendum. Available at *SSRN*: https://ssrn.com/abstract=2798311 [Consulted 20 July 2021].

Jobin, Anna, Ienca, Marcello & Vayena, Effy (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389-399.

Johnson, Deborah. (1996). Forbidden knowledge and science as professional activity. *The Monist*, 79(2), 197-217.

Johnson, Deborah (1999). Reframing the question of forbidden knowledge for modern science. *Science and Engineering Ethics*, 5(4), 445-461.

Jordan-Young, Rebecca (2011). *Brain storm: The flaws in the science of sex differences*. Harvard University Press.

Kelly, Liz (1987). The Continuum of Sexual Violence. In Hanmer, Jalna & Maynard, Mary (eds.). *Women, Violence and Social Control* (46-60). London: Palgrave Macmillan.

Kempner, Joanna, Perlis, Clifford & Merz, Jon (2005). Forbidden knowledge. *Science*, 307(5711), 854.

Khan, Saan, Xiaoxuan, Liu, Siddharth, Nath & et al. (2021). A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *The Lancet Digital Health*, 3(1), 51-66.

Kikerpill, Kristjan (2020). Choose your stars and studs: the rise of deepfake designer porn, *Porn Studies*, 7(4), 352-356.

Kimmel, Michael (2000). *The gendered society*. New York: Oxford University Press.

Kuhn, Thomas (1971). *La estructura de las revoluciones científicas*. México: Fondo de Cultura Económica.

Kusters, Remy, Misevic, Dusan, Berry, Hugues & et al. (2020). Interdisciplinary Research in Artificial Intelligence: Challenges and Opportunities. *Frontiers in Big Data*,3,45.

Latonero, Mark (2018). Governing artificial intelligence: Upholding human rights & dignity. *Data&Society*, 1-37.

Le, Thai Hoang (2011). Applying Artificial Neural Networks for Face Recognition. *Advances in Artificial Neural Systems*, 1687-7594.

Lee, Suk Kyeong (2018). Sex as an important biological variable in biomedical research. *BMB reports*, 51(4), 167-173.

Maddocks, Sophie (2020). 'A Deepfake Porn Plot Intended to Silence Me': exploring continuities between pornographic and 'political' deep fakes. *Porn Studies*, 7(4), 415-423.

Malane, Rachel (2005). *Sex in Mind: The Gendered Brain in Nineteenth-Century Literature and Mental Sciences*. New York: Peter Lang.

McGlynn, Clare & Rackley, Erika (2017). Image-based sexual abuse. *Oxford Journal of Legal Studies*, 37(3), 534-561.

McGlynn, Clare, Johnson, Kelly, Rackley, Erika, Henry, Nicola, Gavey, Nicola, Flynn, Asher & Powell, Anastasia (2021). 'It's Torture for the Soul': The Harms of Image-Based Sexual Abuse. *Social & Legal Studies*, 30(4), 541-562.

Milton, Kay (1979). Male Bias in Anthropology. *Man*, 14(1), 40-54.

Nissani, Moti (1997). Ten cheers for interdisciplinarity: The case for interdisciplinary knowledge and research. *The Social Science Journal*, 34(2), 201-216.

Pastor-Galindo, Javier, Zago, Mattia, Martínez, Gregorio et al. (2020). Spotting political social bots in Twitter: A use case of the 2019 Spanish general election. *IEEE Transactions on Network and Service Management*, 17(4), 2156-2170.

Powell, Anastasia, Scott, Adrian, Flynn, Asher & Henry, Nicola. (2020). Image-based sexual abuse: An international study of victims and perpetrators. A Summary Report. *Criminology*, 1-15.

Rackley, Erika, McGlynn, Clare, Johnson, Kelly, Henry, Nicola et al. (2021). Seeking justice and redress for victim-survivors of image-based sexual abuse. *Feminist Legal Studies,* 293-322.

Ramón Mendos, Lucas, Botha, Kellyn, Carrano, Rafael et al. (2020). *State-Sponsored Homophobia 2020: Global Legislation Overview Update*. Geneva: ILGA World.

Reiter, Rayna (2012). *Toward an anthropology of women*. London: Monthly Review Press.

Rippon, Gina. (2019). *The gendered brain*. London: The Vodley Head.

Robinson, Melody (2019). Biphobia, Rape Myth Acceptance, and Victim Blame for Bisexual Survivors of Sexual Assault. *OSR Journal of Student Research*, 5(329).

Rolnick, David, Donti, Priya, Kaack, Lynn, Kochanski, Kelly & et al. (2019). Tackling Climate Change with Machine Learning,

arXiv:1906.05433 [cs, stat] [Preprint]. Available at: http://arxiv.org/abs/1906.05433 (Accessed: 7 September 2021).

Russell, Diana (1990). *Rape in marriage*. Bloomington: Indiana University Press.

Sarewitz, Daniel (2016). The pressure to publish pushes down quality. *Nature*, 533(7602), 147.

Smith, David (1978). Scientific Knowledge and Forbidden Truths. *The Hastings Center Report*, 8(6), 30-35.

Tajalli, Payman (2021). AI ethics and the banality of evil. *Ethics and Information Technology*, 447-454.

Viseu, Ana (2015). Integration of social science into research is crucial. *Nature*, 525(7569), 291.

Walby, Sylvia (1992). *Theorizing patriarchy*. Oxford: Blackwell.

Wang, Yilun & Kosinski, Michal (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, 114(2), 246-257.

Welzer-Lang, Daniel (2008). Speaking Out Loud About Bisexuality: Biphobia in the Gay and Lesbian Community. *Journal of Bisexuality*, 8(1-2), 81-95.

Westerlund, Mika (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9(11).

Wong, Karen & Dobson, Amy (2019). We're just data: Exploring China's social credit system in relation to digital platform ratings cultures in Westernised democracies. *Global Media and China*, 4(2), 220-232.

World Economic Forum (2018). *The Global Gender Gap Report 2018*.

Young, Erin, Wajcman, Judy & Sprejer, Laila (2021). Where are the Women? Mapping the Gender Job Gap in AI. Policy Briefing: Full Report. *The Alan Turing Institute*.

Youyou, Wu, Kosinski, Michal & Stillwell, David (2015). Computer-based personality judgments are more accurate than those made

by humans. *Proceedings of the National Academy of Sciences*, 112(4), 1036-1040.