Artificial Intelligence and inclusion: an analysis of bias against the poor

Georgina Curto Rex

UNIVERSITAT RAMON LLULL

DOCTORAL THESIS

Title	Artificial Intelligence and inclusion: an analysis of bias against the poor
Presented by	Georgina Curto Rex
Centre	IQS School of Management
Department	Chair of Ethics and Christian Thought
Directed by	Flavio Comim

IQS SCHOOL OF MANAGEMENT UNIVERSITY RAMON LLULL

Artificial intelligence and inclusion: an analysis of bias against the poor

Thesis presented by: Georgina Curto
Director of thesis: Dr. Flavio Comim

Thesis presented to obtain the title of Doctorate degree by the University Ramon Llull
PhD Program: Business and Territorial Competitiveness, Innovation and Sustainability (CETIS)

Barcelona, April 2022

Acknowledgements

This doctoral thesis has been a long journey and a personal commitment. During the adoption of my first child, Josef, I spent three years living among Spain and Morocco, surrounded by a cheerful crowd of children in the Crèche des Enfants Abandonnés Hopital Hassan II d'Agadir, where I learnt the effect that stigma has on a human being, especially when suffering it from a very young age. I want to thank all the children in the Crèche, and my son Josef in particular, for teaching me how to live with joy and dignity even when being discriminated from the very beginning of your life.

A long journey requires a good guide and Dr. Flavio Comim has been an exceptional Thesis Director, an endless and generous source of knowledge and a patient professor who has always been by my side, listening and supporting my own research interests while showing me the way to materialise them into this academic work.

I would like to thank Dr. Enrique Navarro, my life companion during this PhD, for always encouraging me and providing me with invaluable support, sometimes in extremely difficult circumstances, to become a researcher.

Finally, I would like to thank both my children, Josef and Zoe, for making such a big effort to share their mum with the academic research. I dedicate this thesis to them.

The study leading to chapter 4 of this thesis was partially financed by the Aristos Campus Mundus, a joint program of the Ramon Llull, Deusto and Comillas universities.

Summary

Philosopher Adela Cortina coined the term aporophobia in 2017 to describe why wealthy foreigners are welcome while those that are poor are ignored, rejected and even suffer from verbal and physical attacks. While the eradication of poverty is the first Sustainable Development Goal of the United Nations 2030 Agenda, this discriminatory practice has gone unnoticed despite constituting a brake for the success of the policies aimed at mitigating poverty and having devastating effects for the dignity of the persons affected. Additionally, aporophobia constitutes an aggravating factor for other historically identified kinds of discrimination in terms of gender, ethnic and sexual orientation.

In our days, discrimination occurs both in the digital and the tangible world, reinforcing one another in an "onlife" reality. Discriminating the poor in the digital world, however, has some aggravating factors since, among other reasons, there is an overestimation of the information provided by artificial intelligence (AI) models, which in fact are trained on historical data provided by the users online behavioural data. While AI models replicate, reinforce and often aggravate the discrimination patterns existing in society, AI providers do not publicly acknowledge the existence of bias in their models and it is difficult to define accountability when AI systems learn from millions of data obtained from anonymised users. In this context, the European Commission's Ethics Guidelines for Trustworthy AI are part of a framework that seeks to regulate basic agreed human rights in the online environment. However, AI practitioners do not know how to apply conceptual principles such as "diversity, non-discrimination and fairness, including the avoidance of unfair bias" in practice when programming the actual AI models. On the

other hand, the literature aiming to tackle the discrimination topic in AI, does it from an algorithmic point of view and in controlled environments, without providing a strong ethical framework that clarifies neither the nature, the causes nor a way to effectively and realistically deal with discrimination in the onlife. The need for academic analysis to support the identification and mitigation of aporophobia is even more urgent, since the poor are not considered a historically discriminated group in the EU regulatory framework for Trustworthy AI and poverty is not described as a "sensitive attribute" in AI literature.

In this context, this doctoral thesis seeks to explain, provide evidence and mitigate the phenomenon of aporophobia in the onlife. The structure of this doctoral thesis consists on the following chapters: chapters 1 and 7 constitute the introduction and conclusions of the thesis; chapter 2 provides a conceptual framework for aporophobia, identifying the circumstances of this discriminatory practice; chapter 3 analyses the relativeness of the perception of AI fairness, how capitalism social recognition order constitutes an aggravator for aporophobia by adding an element of blame for being poor and how aporophobia is translated into the AI environment; chapter 4 provides empirical evidence of the existence of aporophobia in the social networks and also in AI Natural Language Processing models that are used to develop apps in critical sectors such as health, education and justice; chapter 5 provides a hands-on Artificial Intelligence Bias Mitigation Process (AIBMP) that seeks to apply the Trustworthy AI principle of "diversity, non-discrimination and fairness, including the avoidance of unfair bias" by proposing specific pro-ethical actions within each step of the AI models' development process; finally chapter 6 presents one of the lines of future research, namely using an AI norm optimisation approach to generate simulations that allow to foresee how

aporophobia actually affects poverty levels, providing insights that could guide a new generation of poverty reduction policies, acting not only on redistribution but also on discriminatory issues.

Key words: aporophobia, discrimination, bias, artificial intelligence, poverty

Resumen

La filósofa Adela Cortina acuñó el término aporofobia en 2017 para describir porqué los extranjeros ricos son bienvenidos mientras los pobres son ignorados, rechazados e incluso sufren ataques verbales y físicos. Mientras la erradicación de la pobreza es el primer Objetivo de Desarrollo Sostenible de las Naciones Unidas en la Agenda 2030, esta práctica discriminatoria ha pasado desapercibida a pesar de constituir un freno para el éxito de las políticas destinadas a mitigar la pobreza y tener efectos devastadores para la dignidad de las personas afectadas.

Actualmente, la discriminación se produce tanto en el ámbito tangible como en el digital, reforzando uno al otro en la realidad "onlife". La discriminación de los pobres en el ámbito digital va acompañada de factores agravantes porque, entre otros motivos, existe una sobrevaloración de la información proporcionada por los modelos de inteligencia artificial (IA), que sin embargo están entrenados a partir de datos históricos proporcionados por el comportamiento de los usuarios online. Mientras los modelos de IA replican, refuerzan y a menudo agravan los patrones de discriminación existentes en la sociedad, los proveedores de IA no reconocen abiertamente la existencia de sesgos en sus modelos y resulta difícil asignar responsabilidades cuando los sistemas de IA aprenden a partir de millones de datos obtenidos de usuarios anonimizados. En este contexto, las Directrices Éticas para IA Fiable de la Comisión Europea son parte de un marco regulador para los derechos humanos básicos en el entorno online. Pero el personal técnico especializado IA no tienen las herramientas necesarias para aplicar principios conceptuales como "diversidad, no discriminación y justicia, incluyendo la prevención del sesgo injusto" en la práctica cuando están programando los modelos de IA. Por otro

lado, la bibliografía en el ámbito de discriminación y IA trata la temática desde el punto de vista de los algoritmos y en un entorno controlado, sin apoyarse en un marco ético robusto que clarifique la naturaleza y las causas del fenómeno así como la forma de lidiar con la discriminación de manera efectiva y realista en el onlife. La necesidad de un análisis académico sobre cómo identificar y mitigar el sesgo es incluso más urgente cuando la pobreza es el motivo de discriminación, ya que la aporofobia no se consideran un grupo históricamente discriminado en el marco europeo para una IA Fiable ni como un "atributo sensible" en la literatura de IA.

En este contexto, esta tesis doctoral tiene el propósito de explicar, proporcionar evidencia empírica y mitigar el fenómeno de la aporofobia en el onlife. La estructura de la tesis doctoral consta de los siguientes capítulos: los capítulos 1 y 6 constituyen la introducción y conclusiones de la tesis; el capítulo 2 proporciona un marco conceptual de la aporofobia, identificando las circunstancias de esta práctica discriminatoria, el capítulo 3 analiza la relatividad de la percepción de la justicia en IA, explica cómo el orden de reconocimiento social del capitalismo constituye un agravante de la aporofobia, añadiendo el elemento de culpa por el hecho de ser pobre, y cómo la aporofobia se traslada al ámbito de IA; el capítulo 4 proporciona evidencia empírica sobre la existencia de la aporofobia en las redes sociales y también en los modelos de Procesamiento de Lenguaje Natural en IA que se utilizan para desarrollar aplicaciones digitales en sectores tan críticos como son la sanidad, la educación y la justicia; el capítulo 5 proporciona un Proceso de Mitigación del Sesgo en Inteligencia Artificial (PMSIA) que tiene el propósito de aplicar el principio de IA Fiable de "diversidad, no discriminación y justicia, incluyendo la prevención del sesgo injusto" proponiendo acciones pro-éticas durante cada paso del proceso de desarrollo de los modelos de IA; por último el capítulo 6 presenta una de las futuras líneas

de investigación, concretamente un enfoque normativo de optimización mediante IA para generar simulaciones que permitan prever cómo la aporofobia afecta los niveles de pobreza, proporcionando información que podría guiar una nueva generación de políticas contra la pobreza, actuando no sólo a nivel redistributivo sino también en el ámbito de la discriminación.

Palabras clave: aporofobia, discriminación, sesgo, inteligencia artificial, pobreza

Resum

La filòsofa Adela Cortina va encunyar el terme aporofòbia en 2017 per a descriure per què els estrangers rics són benvinguts mentre que els pobres són ignorats, rebutjats i fins i tot pateixen atacs verbals i físics. Mentre la eradicació de la pobresa és el primer dels Objectius de Desenvolupament Sostenible de les Nacions Unides en l'Agenda 2030, aquesta pràctica discriminatòria ha passat desapercebuda malgrat que constitueix un fre per a l'èxit de les polítiques destinades a mitigar la pobresa i que té efectes devastadors per a la dignitat de les persones afectades.

Actualment, la discriminació es produeix tant en l'àmbit tangible com en el digital; l'un reforça l'altre en la realitat "onlife". La discriminació dels pobres en l'àmbit digital va acompanyada de factors agreujants perquè, entre altres motius, existeix una sobrevaloració de la informació proporcionada pels models d'intel·ligència artificial (IA), que tanmateix estan entrenats a partir de les dades històriques proporcionades pel comportament dels usuaris online. Mentre els models d'IA repliquen, reforcen i sovint agreugen els patrons de discriminació existents a la societat, els proveïdors d'IA no reconeixen obertament l'existència de biaixos en els seus models i resulta difícil assignar responsabilitats quan els sistemes d'IA aprenen a partir de milions de dades obtingudes d'usuaris anonimitzats. En aquest context, les Directrius Ètiques per a una IA Fiable de la Comissió Europea són part d'un marc regulador per als drets humans bàsics en l'entorn online. Però el personal tècnic especialitzat en IA no té les eines necessàries per a aplicar els principis conceptuals de "diversitat, no discriminació i justícia, incloent la prevenció del biaix injust" en la pràctica quan estan programant els models de IA. D'altra banda, la bibliografia en l'àmbit de la discriminació i IA tracta la temàtica des del punt de vista dels

algoritmes y en un entorn controlat, sense recolzar-se en un marc teòric robust que clarifiqui la naturalesa i causes del fenomen i com lidiar amb la discriminació de manera efectiva i realista en el onlife. La necessitat d'un anàlisi acadèmic sobre com identificar i mitigar el biaix és fins i tot més urgent quan la pobresa és el motiu de discriminació, ja que l'aporofòbia no es considera un grup històricament discriminat en el marc europeu per a una IA Fiable ni un "atribut sensible" en la literatura d'IA.

En aquest context, aquesta tesi doctoral té el propòsit d'explicar, proporcionar evidència empírica i mitigar el fenomen de l'aporofòbia en el onlife. L'estructura de la tesi doctoral consta dels següents capítols: els capítols 1 i 6 constitueixen la introducció i conclusions de la tesi; el capítol 2 proporciona un marc conceptual de l'aporofòbia, identificant les circumstàncies d'aquesta pràctica discriminatòria; el capítol 3 analitza la relativitat de la percepció de la justícia en IA explica com l'ordre de reconeixement social del capitalisme constitueix un agreujant de l'aporofòbia, afegint l'element de culpa pel fet de ser pobre, i de cóm l'aporofòbia es trasllada a l'àmbit de la IA; el capítol 4 proporciona evidència empírica sobre l'existència de l'apofobòbia en les xarxes social i també en els models de Processament de Llenguatge Natural en IA que s'utilitzen per a desenvolupar aplicacions digitals en sectors tan crítics com són els serveis de salut, educació i justícia; el capítol 5 proporciona un Procés de Mitigació del Biaix en la Intel·ligència Artificial (PMBIA) que té el propòsit d'aplicar el principi d'IA Fiable de "diversitat, no discriminació i justícia, incloent la prevenció del biaix injust" proposant accions pro-ètiques durant cada pas del procés de desenvolupament dels models de IA; per últim, el capítol 6 presenta una de les futures línies de recerca, concretament un enfocament normatiu d'optimització mitjançant IA per a generar simulacions que permetin preveure com l'aporofòbia afecta els nivells de pobresa, proporcionant informació que podria guiar una nova generació de

polítiques contra la pobresa, actuant no només a nivell redistributiu sinó també en l'àmbit de la discriminació.

Paraules clau: aporofòbia, discriminació, biaix, intel·ligència artificial, pobresa

Table of contents

Chapter 1. Introduction

The study of the bias against the poor in the digital world brings together two key issues of our times: the role of wealth as an indicator of social success and the use of behavioural data as a fuel for economic activity; both are part of the entangled dilemmas of knowledge, authority and power of the capitalist information civilization (Zuboff 2019). While the poor were already discriminated in the stratified society of the old feudal regime, where social status was decided at birth (Piketty 2014), industrial capitalism aggravates this discriminatory phenomenon because the "individualistic achievement principle" (Fraser and Honneth 2003) emerges as a new criterion of social-steam, making the poor often considered responsible for their fate. In welfare states, the rhetoric of equal opportunity and the tyranny of merit described by Sandel (2020b). exacerbates even more the shame suffered by the poor. Currently, the so-called "AI capitalism" (Coeckelberg 2022) or "surveillance capitalism" (Zuboff 2019) contribute to the discrimination against the poor in a more efficient and opaque manner. The chapters that conform this thesis aim to, first of all, offer an analysis of the nature of bias, with specific focus on bias against the poor in the "onlife", using Floridi's expression (2015), informed on a multidisciplinary research on social economics, psychology, political philosophy, sociology, ethics and business analysis. This thesis also provides empirical evidence about bias against the poor and intends to shed some light on the ways forward to tackle the issue, by proposing an inclusive AI development process, offering some preliminary ideas for acting on the AI industry value chain and proposing alternative business models that openly increase the perception of AI fairness. A final objective is to present forthcoming research derived from this thesis which seeks to inform a new generation of poverty reduction policies by acting on aporophobia (rejection of the poor). .

Chapter 2 of the thesis seeks to provide a conceptual background to understand the diversity of biological, economic, political, social, cultural and psychological circumstances that explain this

phenomenon, which do not act in isolation, rather feed one another creating a complex construct that is difficult to combat. Philosopher Adela Cortina coined the term aporophobia and described it in a context of contractualism, explaining that our brain is aporophobic as a result of our biological programming for mutual help (2017). In the capitalist recognition order, aporophobia is aggravated because the poor are blamed for their fate and considered undeserving help (Arneson 1997; Everatt 2009; Nunn and Biressi 2009). This is incorporated in our collective moral framework and goes unnoticed since aporophobia is part of our beliefs are therefore of the way we interpret reality (Ortega Gasset 1942). At a micro-personal level, this construct has an impact on the psychological well-being of the poor, making it more difficult to overcome the financial difficulties. At macro-international and meso-national levels, the poor are considered undeserving to receive help, which is translated into more restrictive welfare policies (Applebaum 2001).

The existing vicious circle of aporophobia and poverty achieves an even higher level of complexity and speed in the online world. AI is not neutral in terms of values, it incorporates a morality (Ausín and Robles Carrillo 2021) and it influences users' behaviours. AI models are trained on Big Data from social networks, reinforcing and even magnifying prejudices, discrimination, stereotypes and bias existing in society. The overestimation of AI, which in fact currently lacks of common sense (Mántaras 2017), and the lack of transparency offered by digital tools aggravate the problem, since there is little questioning on the information they provide (O'Neal 2016; Fry 2018). Moreover, individuals behavioural data collected through a variety of devices is used to implement individually-tailored digital services (Zuboff 2019) which, once more, segregate between the rich and the poor (Eubanks 2018). Since the acquisition of knowledge and critical thought are antidotes to prejudices and the resulting discriminative actions (Allport 1954), the fact that online discrimination occurs within a black box (von Eschenbach 2021) creates especial damage because it does not allow for critical examination.

The articles that conform chapters 3, 4 and 5 of this doctoral thesis are centred on inclusion and bias against the poor in Artificial Intelligence (AI). Chapter 3 analyses the specificities of aporophobia in AI, where the poor are often excluded from health services or job interviews as a result of what has been called the "scored-society" (Benjamin 2019) or the "digital poorhouses" (Eubanks 2018). The chapter also aims to translate how the different contextual perceptions of fairness can be incorporated in AI and explain the changes in the social recognition orders resulting from the value chain and business models of the so-called "AI capitalism" (Coeckelbergh 2022) or "surveillance capitalism" (Zuboff 2019). Chapter 4 provides empirical evidence of the phenomenon of aporophobia by measuring bias against the poor in pretrained Google Word2vec, Twitter and Wikipedia GloVe word embeddings, using vector world representations, a state-of-the-art technique applied to identify online stereotypes regarding other historically discriminated groups (Bolukbasi et al. 2016; Manzini et al. 2019; Nadeem et al. 2020). Chapter 4, therefore, constitutes an example where AI helps us identify and monitor existing discrimination in society, since word embeddings are trained on historical big data obtained from Google News, Wikipedia and Twitter. Chapter 5 is dedicated to the mitigation of bias in AI models from the design stage and through the participation of stakeholders, including users that belong to historically discriminated groups. While a legal framework is being implemented in the European Union recognising the principle of "diversity, non-discrimination and fairness, including the avoidance of unfair bias" (European Comission 2021), AI practitioners have difficulty to comply with this principle in practice. There is a growing number of articles that identify this gap and urge for procedures and guidelines to translate Trustworthy AI principles into practice (Ibáñez and Olmeda 2021; Morley et al. 2021b). Although there is a great deal of literature that describe algorithmic methods to debias AI models (Bolukbasi et al. 2016; Zhao et al. 2018; Manzini et al. 2019; Nadeem et al. 2020) or even incipient methods to build AI models according to pre-defined values (Jiang et al. 2021), these approaches only work in controlled environments. On the other hand, it is not possible to draw a hard line between what is sufficient evidence of bias in absolute terms, since it is based on our values and AI models should be compatible with the societies they operate (Carman & Rosman, 2020). Capter 5 of this doctoral

thesis provides a hands-on pro-ethical design process that aims to support AI development teams to identify, mitigate and monitor bias in practice, in specific cultural frameworks, when using real data. Chapter 6 describes a future line of research in which AI is used as a tool to work towards the achievement of the first of the United Nations Sustainable Development Goals (eradicate poverty). The purpose of this particular study is to provide evidence whether lower levels of bias against the poor would contribute to a decrease in the actual poverty levels. By using an AI agent-based social simulation, this research line aims to provide useful data for a completely new path for poverty reduction policies, based not only on redistribution but also on public awareness to mitigate discrimination against the poor. Finally, the conclusions of this thesis provide a first glance of other future lines of research that derive from the presented work.

This doctoral thesis presents a first approach to explain, provide evidence and mitigate bias against the poor both online and offline. The chapters of the thesis are independent essays, two of which were created as a result of multidisciplinary work in collaboration with AI technical researchers. In particular, chapter 4 is the result of a collaboration with Dr. Mario Fernando Jojoa Acosta and Dr. Begoña García-Zapirain (eVida Research Laboratory of the University of Deusto). Chapter 6 was created with the participation of Nieves Montes, Dr. Nardine Osman and Dr. Carles Sierra (Institut d'Investigació en Intel.ligència Artificial (IIIA – CSIC)).

Chapter 2. Aporophobia. The unnoticed barrier for poverty reduction

Summary of the chapter

This paper explains how aporophobia, the rejection of the poor, affects political decision-making and economic outcomes. It describes why this phenomenon has gone unnoticed despite its effects at a macro-global, meso-national and micro-personal levels. It explores the relationship of the triad aporophobia, poverty and inequality. It identifies the diversity of biological, economic, political, social, cultural and psychological circumstances for this phenomenon, describing the strain between the different motivations and the values behind this discriminatory practice. The paper answers why aporophobia constitutes a brake for poverty reduction, which in itself, constitutes an instrumental reason to identify and mitigate this phenomenon. Additionally, this study explores the psychological and philosophical elements of aporophobia, since this discriminatory practice should also be mitigated for an intrinsic reason: the dignity of the persons affected.

Key words: aporophobia, poverty, inequality, discrimination.

2.1. Introduction

According to Cortina (2017), one of the reasons why some foreigners are rejected while some others are welcome has to do with the fact that some are affluent while others are poor. By creating the term “aporophobia” Cortina has taken the first step to denounce this scourge, since a phenomenon that does not have a name can be more easily ignored even though the realities that it addresses are appalling. Cortina explains that Spain received more than 65 million foreign tourists in 2016, who are clearly welcomed. However, the same year Europe was reluctant to open the door to 160.000 refugees, as recently agreed by member countries, despite the 5 million euros from the European Commission during 2 years for the inclusion of these people. Unfortunately, these numbers are not that high if we consider that 79,5 million people in the world live far from

their homes due to war, violence or serious violation of their fundamental rights, according to ACNUR. At the end of 2019, 32,8 million people were looking for shelter in other countries, being 68% of them from Syria, Venezuela, Afghanistan, South Sudan and Myanmar, according to CEAR (Spanish Commission for helping the refugees). As Cortina states: "we are not disturbed by foreigners. We are disturbed because they are poor" (2017: 14).

The motivation of this article, therefore, is to offer a conceptual analysis that builds on the term coined by Cortina, in order to analyse the circumstances, according to which people reject millions of human beings from different races, ages, gender and cultures that are forced to leave their homes, when it seems that poverty is all that they have in common. The article also aims to explain the conditions that drive people to reject the poor that are part of their own community and we are even ashamed of being poor ourselves. The implications of putting forward a conceptual framework for aporophobia should not be underestimated. The elite perceptions of the poor shapes the values that are behind social institutions, which constitute a powerful obstacle for successful initiatives on poverty reduction (Reis et al. 2005). Therefore, the article aims to characterise how aporophobia could affect the possible solutions to poverty. This question is the connecting threat throughout the article and constitutes an instrumental justification of the study of aporophobia. However, the research on aporophobia is also justified from an intrinsic point of view: for the dignity of the persons affected.

The paper is organized into six parts. The first puts forward a conceptual model to systemize the main features of aporophobia, explaining why this phenomenon has not been described before as a specific kind of discrimination. The second part explores the relationship between aporophobia, poverty and inequality. The third part examines the most important circumstances that contribute to the phenomenon of aporophobia. The fourth part presents the reasons why aporophobia constitutes a brake for poverty reduction, including some considerations for policy making (Table 1). The fifth part describes the feedback between the different circumstances that constitute the construct of aporophobia, creating a vicious circle that aggravates the stigmatization of the poor. Finally, the last part concludes and highlights that discrimination against the poor is integrated

from personal behaviors to macro-economic policy making, explains why it has been tolerated and why this study provides a new path for poverty and inequality reduction.

2.2.The rejection against the poor as a phobia

Since Cortina (2017) describes the rejection against the poor as a phobia, it is worth analysing the nature of phobias as a psychological phenomenon and see how it corresponds to the rejection of the poor. First of all, we must clarify that phobias are psychological pathologies connected to fear, which generate a complex output of relatively independent manifestations into three response systems: verbal-cognitive, behavioural and phycological (Lang 2004). However, while some fears are necessary for humans to survive as species (adaptative fears), phobias are exaggerated and often a kind of disabling fear that do not correspond to certain stimulus (Marks 1969). In the particular case of aporophobia, the fear is originated from prejudices against the poor, which are overgeneralized and erroneous beliefs (Allport 1954). Therefore, we can consider that aporophobia is a social pathology that implies the rejection of individuals as a result of prejudices against the poor, understanding poverty as the lack of freedom to carry out a meaningful life with dignity (Sen 2001; Navarro 2002; Nussbaum 2012; Cortina 2017; Esquembre 2019; Comim et al. 2020).

Thus, aporophobia generates a distorted response based on prejudices, which according to Allport (1954) can be expressed in different degrees of negative action, from antilocution, avoidance and discrimination to physical attach and even extermination. It is therefore logical to wonder if aporophobia is necessary for people at all, either as species or as individuals. According to Morgados (2017), we find that emotions are central to human life, but expressions of hate are unnecessary. In other words, people would be better off without them. In fact, in the study performed by Aumer-Ryan and Hatfield (2007), 30% of participants (N=591) declared that they had never experienced the feeling of hate towards another person in their life. If the extreme expressions of negative action caused by aporophobia are considered expressions of hate (such

as hate speech and hate acts), we can suggest that, from a psychological point of view and at least in these extreme cases, aporophobia could be mitigated.

If phobias are not useful, can we recognize them? In order to answer this question, it is essential to highlight that aporophobia, as it is the case with all rejection attitudes, is founded on our beliefs (Cortina, 2017) and therefore it is latent and part of the framework that sustains our life course. As Ortega explained (1940), we have ideas and we live in beliefs. In other words, it is often hard to see the difference between beliefs and reality itself, since we are not conscious of them and beliefs conform our reality.

How do we identify aporophobic thoughts and actions, then, if we might not even be conscious of them? As shown in Figure 1, we propose a model to recognize aporophobia based on Ortega's (1940) description of the way we interpret reality. According to the model, when experiencing a rejection attitude, first of all people should identify what belief or set of beliefs are triggering that reaction. By doing so, people transform that belief into an idea, since they are conscious of it, and therefore they can evaluate whether it is founded or not on their ethical framework, such as the Universal Declaration of Human Rights. If it is so, the reaction attitude can be completely justified and even necessary, since critical capacity is essential to discern between what is fair and what is unfair (Cortina, 2007). However, they might find out that the rejection attitude is founded on beliefs that do not correspond to their values, in which case it might respond to a personal or social phobia, as it is the case with aporophobia (Fig 1).

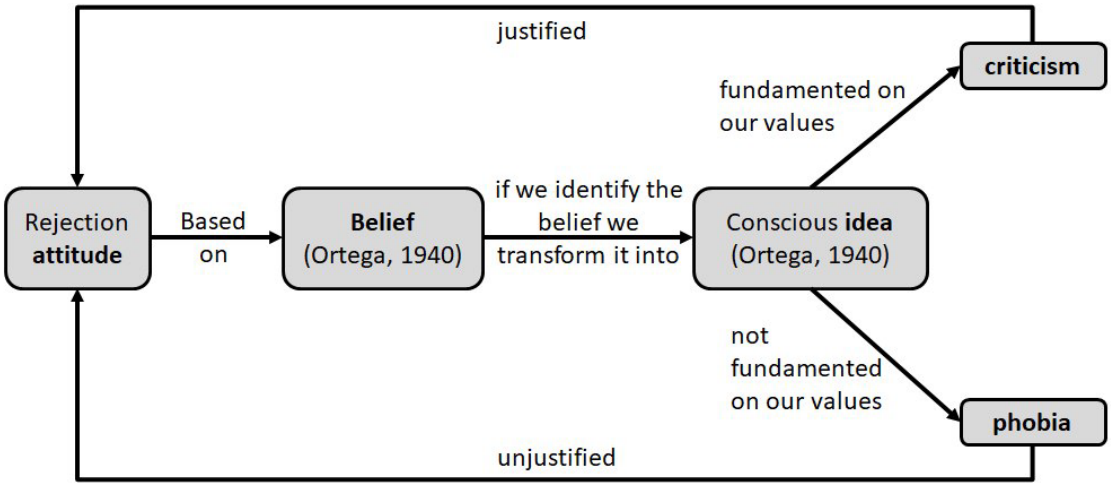

Figure 1. How to identify when oneself is being aporophobic, based on Ortega's (1940) description of the way we interpret reality. Source: author's creation

In other words, not all rejection attitudes are objectionable. Some are legitimate and desirable, such as rejection towards the actions that violate basic human rights. But what are the shared values that can justify criticism and rejection attitudes in our globalized and multicultural context? Cortina acknowledges that we share, as human beings, undeniable common values and moral principles. Therefore, the "minimum ethics" (Cortina 1986) and "cosmopolitan ethics" (Cortina 2021) pose the challenge to discover the shared ethical capital that make us human in a global interconnected context.

2.3.The triad: aporophobia, inequality and poverty

The concept of income inequality is at the root of aporophobia, since there would be no discrimination of the poor if there was no inequality at all. Inequality is therefore the breeding ground for aporophobia. However, a discussion on inequality, poverty (understood as a manifestation of an extreme form of inequality) and aporophobia is complex since these concepts mix materiality and values or, in other words, what it is tangible with what is intangible.

Let's start with the tangible or material aspects of the discussion. In terms of income inequality, several studies have concluded that it affects the pace at which growth enables poverty reduction

(Aghion et al. 1999; Galor and Moav 2004; Ravallion 2004; Ostry et al. 2014). In the current economic context this is particularly relevant, since, according to The World Bank (2020), COVID-19 and the resulting economic crisis are reversing more than two decades of poverty reduction, pushing between 88 and 115 million people into poverty worldwide and threatening to widen income inequalities. But that is not all, since conflict and climate change may force rising numbers of people into poverty in the medium term. In this context, the UN Sustainable Development Goals, which in 2020 included the reduction of inequality based on income in the 2030 Agenda (2020), have become even more pertinent.

Although in the past rising inequality was often seen as a necessary evil, the price paid for growth which would eventually benefit the poorest, today there is a growing consensus backed by research from the International Monetary Fund (Dabla-Norris et al. 2015), among others, that growing inequality implies a brake on growth, in addition to the social costs. Moreover, several studies suggest that, in certain circumstances, income inequality can generate an economic decline (Dabla-Norris et al. 2015).

So, if income inequality matters for growth and poverty reduction, should we not contemplate the hypothesis that aporophobia is a brake for poverty reduction as well? The influence of aporophobia on poverty reduction can be explained at three different levels (Fig 2).

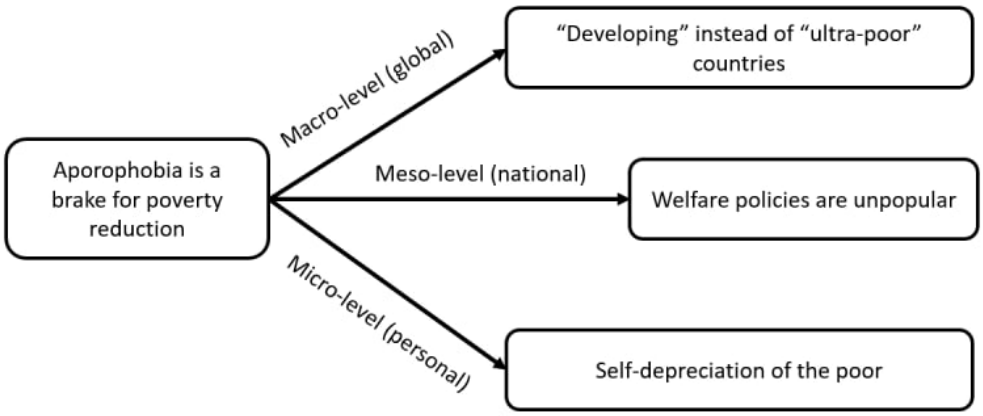

Figure 2. Aporophobia constitutes a brake for poverty reduction at a macro-international, meso-national and micro-personal levels. Source: author's creation

At a macro-international level, the developing countries are considered responsible for their fate, instead of working towards a global equilibrium in areas such as international commerce, cooperation among countries and financial markets (Sampedro, 1972; Tortosa, 2001; Yapa, 2002; Espinosa, 2004; Reis et al., 2005). Sampedro argues that the term "developing" countries is a euphemism which conveys that "progress is underway", when in fact we should be talking about "ultra poor countries" and "marginalised poverty" which "is not the inferior and transitory step of the continuous stairs towards development, but a persistent consequence of development and created as a result of it" (1972 : 20). Sampedro describes "under development" as a much more serious problem than poverty, since "the poor in the traditional world were integrated and felt members of it. The under development adds to the lack of resources the lack of participation" (1972 : 21).

At a meso-national level, the welfare policies are difficult to pass due to the belief that there are poor that deserve this condition (Arneson 1997; Applebaum 2001; Everatt 2009; Nunn and Biressi 2009). Wilson (1996) already cited surveys providing evidence that "whereas a substantial majority of Americans felt that too little was being spent to help the poor, only slightly more than 20 percent in any given year felt that too little was being spent to help those on welfare" (p. 162).

Applebaum also found as a result of the performed surveys that people were more willing to accept liberal policies to aid the poor when the target was from a group perceived to be deserving. For example, if the poor was a person who was made redundant, the survey participants were more prone to help rather than a person perceived not to follow mainstream norms described, for example, as a teenage single mother who did not know who the father was of her child and who refused to take an offered job (Applebaum 2001).

Finally, at a micro-personal level, the self-depreciation of the poor resulting from the social stigma is an additional obstacle to improve their economic situation (Honneth, 1996). Goffman describes the consequences of the stigma suffered by the unemployed, who feel humiliated in their social circles and feel ashamed and insecure (1963). The shame of being poor often becomes a self-fulfilling prophecy.

However, we must bear in mind that the discussion about the relationship between inequality, poverty and aporophobia is more complex, since poverty is not restricted to income. Away from the predominant GDP model, which implies that the quality of life improves in line with the goods and services produced by the country, we find throughout history more human-centered approaches that attempt to improve the standard of living of individuals and deal with poverty and inequality. Sen (2001) builds on this approach of human self-realization already initiated by Adam Smith and Stuart Mill, followed by Rawls, and constructs a conceptual framework where success of a society is evaluated by substantive freedoms that its members enjoy. Sen, therefore, understands poverty as a deprivation of basic capabilities, such as healthcare and primary education. Nussbaum (2012), in turn, defines a specific list of capabilities that individuals should have in order to ensure dignity and equal opportunities. These capabilities are not just abilities residing inside a person, but also the freedom and opportunities created by a combination of personal abilities and the political, social and economic environment.

Therefore, although the inequality of income seems to be the main factor when discussing about poverty, it is only one of the many factors that influence the real opportunities of human beings,

according to the capability approach (Sen 2001). In fact, the scarce income is only an instrumental factor among others such as age, sex, geography, epidemiology or role in the family which also have an influence on individual freedom.

Now, how do aporophobia and inequality relate? We know that in a wealthy country, a higher income is required in order to achieve the same social functioning; what Adam Smith described as the possibility to appear in public without feeling ashamed ([1776] 2020). Therefore, poverty reduction is necessary but not sufficient to eliminate aporophobia, as long as there is inequality (Fig 3).

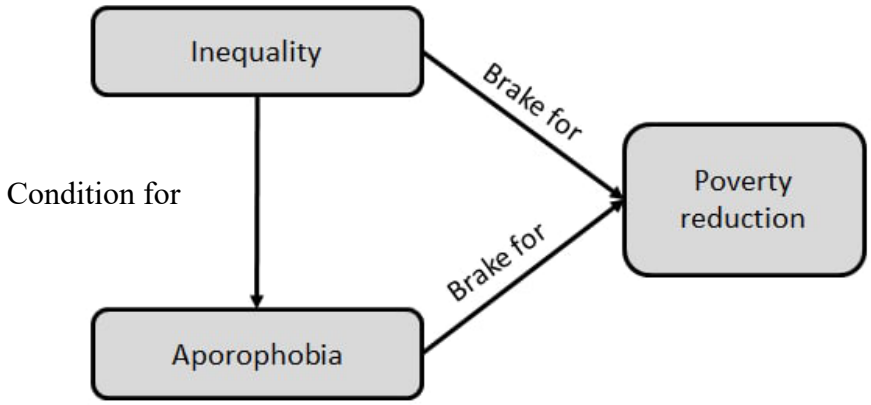

Figure 3. Inequality is a condition for aporophobia. Both inequality and aporophobia constitute a brake for income poverty reduction. Source: author's creation

Let's focus on the intangible aspects of inequality and their relationship with aporophobia, namely, the values behind this concept in a historical context. When we read Aristotle in the IV B.C. (2012), we can be stunned when he describes that the natural order is to have masters and slaves. In 370 B.C, Plato, in turn, states that natural and social inequality are not unfair, but useful in the political community of a polis, since each citizen can be in charge of the tasks he or she is capable of doing (2013). Obviously, when Rousseau ([1798] 1923) defended the concept of equality, which became one of the principles of French Revolution, it was a claim for social justice.

In this sense, the phenomenon of aporophobia can, at first sight, seem particularly striking in a society where equality is one of the theoretical pillars of modernity. Since Rousseau coined this principle, the idea that the state should ensure citizens to be treated equally has enjoyed great success in liberalism and has been included in our legal frameworks. However, as Lyotard (1996) states in his studies on postmodernism, the principle of equality is being questioned. Additionally, equality is a debilitated goal in the multicultural global society we live in, where there is a lack of unitarian referents nor a superior instance capable to generate consensus about what it is that we want to be equal to (Taylor 2009). If the principle of equality is at crisis, we could then ask ourselves why we care about aporophobia at all. Should it not be normal, then, in the framework of liberal capitalist societies to make differences in the treatment of persons according to weather they are rich or poor, as much as it seemed natural for Aristotle to have masters and slaves?

The answer is that aporophobia is of course not acceptable because of the violation of the dignity of human beings that it implies, or in line with Sen's capability approach, aporophobia is intrinsically not acceptable because it limits human agency and freedom. In fact, the principle of equality is at crisis as it was formulated during the French Revolution, but it is not dead at all. Taylor (2009) specifies that equal respect does not necessarily mean being treated equally, since we are all diverse, especially in a multicultural and global environment we live in. What is more, special treatment should be granted to groups that are discriminated, respecting their alterity. In the current context of "cosmopolitan ethics", Cortina changes the focus from equality to justice. Inequality, according to Cortina (2007), does not need to be negative; only when it is unfair. We can therefore specify that this study does not aim to solve inequality but contribute to improve human dignity of the poor. If we take Cortina's (2007) explanation of dignity as "as a link or ligation with oneself: since not only I ask to be respected, not only I have to respect others, but I also have to respect myself" (2007 : chapter 5.6) we can conclude that we ignore the poor, also, because we do not have the dignity to act according to our own principles, because it takes an effort.

To sum up, inequality is at the core nature of aporophobia, both understood as income inequality and in the broader sense, including political freedoms, social facilities, transparency guarantees and protective security, as described in Sen's capability approach (Sen 2001). From an instrumental point of view, both inequality and aporophobia should be reduced since they can constitute a brake for poverty reduction policies. On the other hand, from an intrinsic point of view, both concepts take away the capabilities to fulfill a happy life or, in other words, they take away the person's dignity.

2.4. Circumstances that contribute to aporophobia

Aporophobia is a complex group of phenomena that might be triggered by a wide array of reasons. Without attempting to provide a comprehensive list of these elements, we have grouped them into key categories according to their analytical significance (Fig 4).

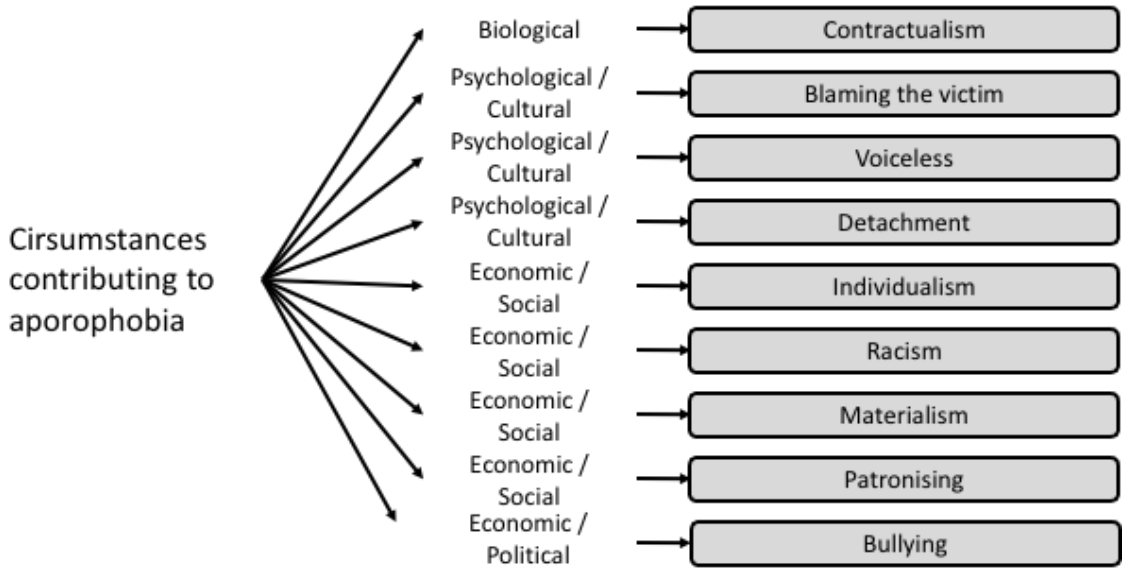

Figure 4. Aporophobia is a phenomenon triggered by a multiplicity of circumstances. Source: author's creation

2.4.1. Contractualism

Cortina states that our brain is aporophobic (2017). Indeed, sometime during the last two million years, our brain incorporated codes based on mutual help, probably as a result of natural selection (Boyd and Richerson 2009), which enhanced social cohesion. The "homo reciprocans" cares about the vulnerable descendants, family and friends but is also suspicious of others unless they have something to offer. Therefore the "aporoi" are the people that have nothing to offer (Cortina 2017).

Darwin ([1871] 2004) explains that altruism is based on help within the group and rejection to strangers, which is one of the keys to understand people's conduct: human beings are willing to give with the expectations to receive something in return. In other words, we are willing to cooperate as a more intelligent strategy to survive rather than through conflict.

Interested cooperation, however, can be based on a diversity of assets other than material wealth (Fig. 5). We can cooperate in exchange of knowledge, care or just by mere socializing, which is essential to human beings. Even in Gary Becker's model of rational allocation, altruism is recognized as a result of getting the sympathy from others (1965). Also Adam Smith recognizes sympathy as part of human nature characterized by mixed motives, where individualism is supported by group membership, but not necessarily only as a result of wealth (Smith [1759] 2016).

Then, can we consider that contractualism is focused on wealth as a result of capitalist recognition order based on material assets? According to Macpherson (2005), current western liberal capitalism has historical roots based on "possessive individualism" that was developed in the framework of the political philosophy of Hobbes and Locke and consolidated as a "market-based possessive society" in the framework of David Hume and Adam Smith. As Macpherson states, the consumer ethos of the advanced contemporary capitalist society has a clear preference on the

human being as “consumer” rather than as “doer and creator”, which can lead to the belief that the person that has no material assets to offer has nothing to offer at all.

2.4.2. Blaming the victim

Adam Smith observes “that wealth and greatness are often regarded with the respect and admiration which are due only to wisdom and virtue; and that the contempt, of which vice and folly are the only proper objects, is often most unjustly bestowed upon poverty and weakness” [1759] 2016 : 69).

Different studies have proved that people differentiate between the deserving poor and the undeserving poor, being the latter, an underclass defined as people with a low income who don’t comply with mainstream norms, with greater number of nonwhites, and individually responsible for their poverty (Applebaum 2001; Nunn and Biressi 2009).

In similar lines, Wilson documents surveys proving that “whereas a substantial majority of Americans felt that too little was being spent to help the poor, only slightly more than 20 percent in any given year felt that too little was being spent to help those on welfare” (1996 : 162). Obviously, this has a clear impact on poverty policies, since they are difficult to pass when it is believed that some poor are to blame for their condition, and an evidence that aporophobia is an obstacle to solve poverty.

This sort of rhetoric has long been acknowledged as ``blaming the victim", meaning that the line of individual and social responsibility for poverty is unclear and therefore, in terms of distributive justice, the poor are in large part to blame. This argument, however, does not exempt the duty to help the ones in need, irrespective of whether they are responsible or not (Arneson 1997; Everatt 2009).

Blaming the poor contributes to reinforce aporophobia, but also the framework of norms and values of our society, as Gans observes (1994). It is well-known that the violations of the norms

of the poor are highly publicized and, as Emile Durkheim pointed out, norm violations "shock collective sentiments" since "wherever a directive power is established, its primary and principal function is to create respect for the beliefs, traditions and collective practices" ([1893] 2014 : 84). Public punishments contribute to preserving and reaffirming the norms. Therefore, by stigmatizing the poor and associating them with stigmatized factors, such as criminality, we may believe that we can avoid economic problems if we behave according to the mainstream norms. Undoubtedly, our beliefs, as defined by Ortega (1940), are involved in blaming the poor when success is understood as the achievement of material wealth in traditional capitalist narrative of competition and a winner-take-all mentality described by Freeman et al (2007) (Fig. 6). The poor, therefore, are believed to be the outsiders of the system, the persons that do not comply with the norms and deserve to be poor; the "losers". In other words, we can say that aporophobia contributes to keeping our liberal economic cultural framework.

Of course blaming the poor has an impact in terms of macro-economy, since people blame, to a great extent, the poor countries for being poor, instead of working towards a global governance that promotes an international equilibrium in areas such as international commerce, cooperation among countries and the financial markets, as suggested by Cortina (2017). That is why Sampedro (1972) states that being underdeveloped is not a transitory step of the growing line but, as described by Bolívar Echevarría (2011), a persistent consequence of development.

The objective of global governance, however, is not to expand the culture and standard of living of developed countries to the rest of the world. We are not talking about material development only. Understanding development as material wealth is a result of a specific culture: the modern technical civilization of western countries as opposed to what Sampedro calls humanized development (1972), which is only feasible if developed countries recognize that there is aporophobia at a macro-international level and are open to a real cosmopolite society as mentioned by Cortina (2017).

2.4.3. Voicesless

Bearing in mind the abundant literature on poverty, studies where the poor are given a voice have not received the recognition they deserve. In fact, we could argue that this acknowledgment on its own can be seen as evidence of aporophobia within the poverty literature. An exception of this lack of representation is the project "Voices of the Poor" (Narayan and Petesch 2002), a series of publications that present poor people's own voices through participatory and qualitative research methods, documenting the testimonies of poor persons and extracting some patterns that these individuals from 23 different countries have in common.

Not surprisingly, a common concern among the testimonies that have been reported by the publication is the widespread social disapproval the poor must face. As Narayan and Petesch (2002 : 30) put it, "the mere fact of being poor is cause of being isolated, left out, looked down upon, alienated, pushed aside and ignored. The ostracism and voicelessness tie together the poor people's experiences across different countries". The project participants perceive this discrimination not only by their co-citizens, but also by the state and private sector institutions. This is particularly shocking in an era characterized by the development of communication technology. But, obviously, the difficulty we are facing is not related to the means of communication, but to what people have to say to each other (Burkett 2000).

Mutual recognition is the nucleus of social life, according to discourse ethics (Honneth 1996; Cortina 2007). The genesis of human mind is not monologic (we do not achieve to be ourselves by our own), but dialogic (Taylor 2009). In other words, recognizing the alterity of others helps one to recognize oneself and the lack of recognition towards the poor implies to live as if they did not exist, as it is acknowledged in the Voices of the Poor (Narayan and Petesch 2002), where "feeling invisible" is described as one of the worse consequences of being poor. As Cortina states, "the poor is precisely, by his or her essence, the potential interlocutor who will never be for real"

(1991: 127) and lack of communication implies the lack of capacity to see and share the suffering and happiness of others (Cortina 2007).

Indeed, our development as human beings depends, also, on the recognition that we obtain from others (Honneth 1996). Therefore, the rejection of the poor can be perceived as such an oppression that is internalized by the affected individuals (Taylor 2009). We could argue, therefore, that aporophobia constitutes a brake on poverty reduction also at a personal level, since the poor often end up feeling "second-class" citizens with no legitimate right to improve their situation. As Taylor describes it, "self-depreciation becomes one of the most powerful instruments of self-oppression" (2009 : 54) since recognition is a vital human need, and it is not simple to liberate oneself from a destructive and imposed identity.

Dealing with the consequences of aporophobia at a personal level is, therefore, a complex issue. According to Honneth (2005), esteem is accorded on the basis of an individual's contribution to a shared project, the elimination of demeaning cultural images of minorities does not provide esteem directly, establishes the conditions under which members of those groups can build self-esteem by contributing to the community.

2.4.4. Detachment

Poverty is often seen as an inevitable flaw of the system that prioritizes continuity of the economic cycle and where citizens have very limited voice. According to Habermas (1990), the "system" has colonized the "world of life" in a way that bureaucratized and impersonal techniques, power and money are introduced so deeply, that individuals feel chained, restricted in their autonomy. Cortina (2007) also acknowledges the incapability to be your own master in politics when you are a servant in economy. As Habermas describes, in this "world of life", values and ideals such as the Declaration of Human Rights do not intend to be reality, since they are always subordinated to the continuity of the economic system. In addition, in our model of democracy, citizens have

little participation in fundamental decisions since there is a lack of real deep discussion, being the mass media the main channel.

To aggravate the situation, there is the collective believe that economy does not have ethical principles (Cortina, 2017). It seems that the "system" with its technical and functional resources might have acquired life of its own and the poor are collateral damages we have to live with. In this context, where poverty is considered a collateral damage, aporophobia is also tolerated.

Kant states ([1789] 2015 : 12): "Inclination is blind and slavish, whether it be of a good sort or not, and, when morality is in question, reason must not play the part merely of guardian to inclination, but disregarding it altogether must attend simply to its own interest as pure practical reason. This very feeling of compassion and tender sympathy, if it precedes the deliberation on the question of duty and becomes a determining principle, is even annoying to right thinking persons, brings their deliberate maxims into confusion, and makes them wish to be delivered from it and to be subject to lawgiving reason alone". As Cortina explains (2007), Kant has inherited the stoic tradition, which opposes the logic of the mind to the logic of the heart and has generated painful outcomes in our society; aporophobia is an example. It is important to point out that Cortina (2007) builds on Kant's argumentations on dignity incorporating the moral drive of compassion. However, Cortina acknowledges that people do not desire to inspire compassion, since it implies a feeling of pity which has patronizing connotations. We can therefore conclude, therefore that compassion is an undesired antidote for aporophobia by both the poor and the non-poor.

2.4.5. Individualism

Bolivar Echevarría (2011) describes individualism as a phenomenon that characterizes modernity, as opposed to the ancestral tradition of communitarianism, where the atom of society was not the individual but a community such as a family. Emile Durkheim ([1893] 2014) acknowledges that the lack of unitarian referents with normative character nor a superior instance capable to generate

consensus in a global multicultural complex society results into a lack of social network. Habermas (2007) also describes our society as a secular, technical-scientific environment, focused on results, where there is no superior instance that generates a social network. Even in the capability approach, the concept of individual agency is at the center of the approach (Sen 2001)

Individualism together with the winner-takes-all traditional capitalist narrative (Freeman et al. 2007), contribute to an "individualistic achievement principle" (Fraser and Honneth 2003 : 147) associated to wealth, which are at the core of aporophobia.

Individualism is exacerbated by the difficulty to connect with persons coming from another culture, since there are no shared values, referents, background, which are crucial to communication. The success of the communicative action, therefore, means that the interlocutors recognize each other not only as beings capable to discuss and follow logical rules, but as people capable to tune in (Cortina, 2007). As Taylor explains (2009), we are talking about philosophical frontiers in our multicultural global societies, in addition to the physical walls that are unfortunately being built. According to Cortina (2017), we live in a wild individualistic liberalism where there is the need to build an intercultural citizenship in order to recover the social network, with shared values. In other words, we live in a globalization where technology and communication are key, but we need to learn to use it to build new forms of community according to global ethics.

2.4.6. Racism

It is well-known that there is a considerable overlap between poverty and race, in particular in countries that had slavery and in countries with immigration. Alessina and Glaeser (2013) document that rejection of the poor, that is aporophobia, is also more important when minority groups are over-represented among the poor. Within the United States, we find that states with lower share of African-Americans offer more generous welfare benefits. In other words, there is

a correlation between the percentage of minority groups within the poor population and aporophobia.

In particular, Alessina and Glaeser use the fractionalization index defined by (Charles Lewis Taylor and Hudson 1972), which measures the probability that two people drawn at random from the population are from different racial, ethnic, or linguistic groups. Alessina and Glaeser observe that there is a clear relationship between this diversity index and the social spending as a share of GDP. Among the 16 countries with racial fractionalization greater than 40%, the mean share of GDP spent of social services is 2.42% on average. In contrast, the racially homogeneous countries where racial fractionalization accounts for less than 10%, the average level of social spending as a share of GDP equals 12.87%.

Additionally, Alessina and Glaeser describe that racial diversity has a higher impact on the degree of redistribution than ethnic or linguistic factors. In fact, the correlation between the racial fractionalization and the degree of redistribution is 66% on average, while ethnicity shows 43% correlation and linguistic factors 41%. To sum up, the more diverse a society is in terms of race, the less the countries spend in redistribution policies, providing evidence that racism is a component of aporophobia.

2.4.7. Materialism

According to Bolivar Echevarría (2011), the new productivity under the neotechnic era defined by Mumford ([1934] 1997) provides western civilization a historical success which gradually transforms the Roman Christian western civilization into the current capitalist society. Therefore, the mainstream old Europe became the focus that irradiated the capitalist modernity model to the rest of the world, where discovery of new instruments and techniques is no longer an accidental or spontaneous action, but a result of scientific breakthroughs which are still not developed completely.

In this framework of productivist faith, aporophobia does not come as a surprise, since poverty does not correspond to the values of modernity linked to the symbiosis of neotechnic development and capitalism. Money is the reward of the "capitalist achievement principle" (Fraser and Honneth 2003) and therefore a value that defines us as individuals when struggling to define our own identity. As Bolivar Echevarría puts it, the promise of emancipation of individuals that was suggested as a result of the neotechnic era has taken place in the opposite direction. Sandel observes that markets and market values are now part of spheres of life where they do not naturally belong and identifies two problems in a society where almost everything is for sale: the first one is inequality and the second is corruption (2013). In this chapter, we identify a third one: aporophobia. In a completely materialistic society, we could say that there is a commodification of people, who are only seen as what they can offer or produce in the market, and the poor are rejected as a result.

2.4.8. Patronising

Yapa (2002) observes, based on Derrida's work (1978), that western philosophical tradition tends to interpret the world as a duality of poor and non-poor, developing and developed, problem and no-problem. Yapa argues that scientists who study the problem are outside the "poverty sector" and therefore in the "no-problem" group. Therefore, the academia is the competent knowing "subject" studying the poor as the needy "object". Under the basis of the naïve claim that science is value free and objective, the solutions proposed for the poor tend to be a mirror of the path followed by the "no-problem" group. Policies are therefore defined to reach this objective, metrics and rankings are created to monitor them, which often have a counteractive effect, since they remind most part of the population that they are shameful and living in the periphery of the global system.

In addition, as Yapa explains, the path of economic development as we know it is completely unfeasible, considering that the wealthier one fifth of the world's population consume four-fifths of the planet resources. In this context, we ask ourselves if developing countries should aim at

having the GDP growth and consumption habits of the western world. Of course, they are entitled to it in terms of global justice, but should we not find a better way to live globally? If we defend agency in terms of Sen (2001) as reduction of poverty and not only as the provision of primary goods, we could surely imagine more frugal ways to carry out a meaningful live and "an approach to development that is freedom-centered" (Sen, 2001: 24) which largely differ from our current capitalist model.

In other words, most population of the western world could be considered poor under Sen's capability approach, in terms of lack of freedom because of the long working hours, big consumption needs for social status and little time to spend with their families. Most citizens of developed countries do not feel free to pursue their happiness. In this context, we can argue that discriminating the persons that have fewer material resources is part of the patronizing ethnocentric attitude of the western world.

2.4.9. Bullying

When talking about the colonies, Maquiavel ([1532] 2012) explains that they can be easily offended because they are poor and dispersed; thus they cannot do harm. Although it might seem obvious, it is important to highlight that the most basic human rights of the poor can be violated because there is the shared belief that they cannot easily damage the rest of the society they live in, or at least not voluntarily and in an organized way.

Hobbes in Leviathan ([1651] 2018) explains that individuals are fair mainly because they are weak and afraid. In other words, Hobbes theory points out the fact that the powerful, those who feel omnipotent, do not perceive the need to comply with moral obligations. If this principle is extrapolated to countries, it constitutes an explanation why wealthy countries and individuals do not commit themselves to comply with the most basic human rights, since mistakenly they assume that the actions on the poor will have no consequences.

However, the critical current circumstances as a result of the COVID-19 are just one more instance of the fact that the bad living conditions of the poor in terms of housing, food or medical support do affect the rest of the planet. What is more, essential activities for our survival as species, such as agriculture, are mainly being performed by a segment of the population that would be considered poor in terms of income.

2.5. Aporophobia as an obstacle for poverty reduction

The following section describes why aporophobia constitutes a brake for poverty reduction based on the circumstances that contribute to the phenomenon identified in the chapter.

Circumstance that contributes to aporophobia	Description	Why is it a brake for poverty reduction	Topics to consider in policy making
Contractualism	The poor are not interesting in terms of cooperation in a market-based society (Cortina 2017; Esquembre 2019)	Exchange is performed amongst the people / institutions / countries with similar wealth and productive capabilities (Sampedro 1972; Tortosa 2001)	Enrich exchanges providing visibility and value to activities often considered outside the market such as care and knowledge (Folbre 2021).
Blaming the victim	Deserving poor (Arneson 1997; Everatt 2009; Nunn and Biressi 2009). By stigmatizing the poor, people belief that they can avoid economic problems if they behave according to mainstream norms (Durkheim, [1893] 2014), (Gans 1994).	The determination of blame associated to poverty plays a role in welfare policies decisions. If the recipients of aid are considered to be responsible for their poverty, welfare policies are more restrictive (Applebaum 2001).	To facilitate acceptance, prioritise programs aiding broad groups of disadvantaged people focusing on “equality of life chances” rather than “equality of individual opportunities” (Wilson 1987)

Voicelessness	The poor are not given a voice, feel invisible and second-class citizens. (Cortina 1991; Narayan and Petesch 2002)	We develop as human beings from the recognition we obtain from others. (Honneth 1996), Self-depreciation is a powerful tool of repression (Taylor 2009).	Dealing with aporophobia at a personal level: recovering self-esteem through the participation of shared projects, since the elimination of demeaning cultural images does not provide recovery of self-esteem directly (Honneth 1996). Encouragement of community projects.
Detachment	In order to avoid painful compassion Kant ([1789] 2015). By stigmatising the poor people reduce the risk of being hurt or angered (Gans 1994). Poverty is seen as a flaw of the system (Habermas 1990).	If economy is considered not to have ethical principles, it is not at the service of people and does not work towards reducing poverty. (Cortina, 2007)	Equal dignity is a more effective antidote to aporophobia than compassion, which has a patronising implication of pity. (Schopenhauer 1993). (Esquembre 2019). Improving quality of democracy with deeper discussions on poverty (Reis et al. 2005), (Habermas 1990), (Esquembre 2019),
Individualism	As opposed to ancestral tradition of communitarism (Echevarría 2011). Lack of social network (Durkheim [1898] 2014). Multicultural global contexts aggravate individualism (Taylor 2009).	The poor, which often are not dedicated to productive activities due to age, disabilities, lower education, children raising, are an obstacle to achieve individual results. Rejecting or ignoring them is more practical to achieve individual results. (Folbre 2021)	Build intercultural citizenship and new ways of communitarism using technology in order to recover the social network and shared values (Cortina, 2017)

Bullying	The poor are not a pressure group and can be ignored – rejected (Maquiavel [1532] 2012)	The needs of the poor are seldomly met	Define criteria to identify injustice in order to avoid bias in favor of economic lobbies. (Cortina, 2007)
Racism	Social spending is lower in racially heterogeneous countries (Alessina and Glaeser 2013)	Aporophobia added to racism contributes to less social spending (Reis et al. 2005)	Anti-racist policies work towards poverty reduction. (Reis et al. 2005)
Materialism	Productivity is a success shared value in capitalist modernity (Echevarría 2011). Money is a value that defines us as individuals when struggling to define our identity (Fraser and Honneth 2003)	There is a commodification of people, who are seen as what they can offer or produce in the market (Sandel 2013), for which reducing poverty is not a priority.	Agreeing on minimum global ethics, where dignity of human beings is recognized(Cortina 2010). For intrinsic reasons, independently from what they can offer to the market (Sen 2001).
Patronising	Discriminating the persons that have fewer material resources is part of the patronizing ethnocentric attitude of the western world. It is considered that poor countries are uncapable since they are still "developing", without considering the difficulties related to environment, international trade and colonial heritage (Sampedro 1972; Yapa 2002)	Topics such as international fair trade, environment protection and colonial heritage are not dealt with as moral equals (Tortosa 2001; de Espinosa 2004)	Explore alternative ways of development according to Sen's capability approach (Sen 2001) focusing on the reduction of not only income poverty, but also in the improvement of individual and collective agency to pursue a happy life (Sen 2001; Nussbaum 2012; Folbre 2021)

Table 1. Reasons why aporophobia constitutes a brake for poverty reduction based on the causes of aporophobia identified in the chapter.

2.6. Aporophobia is a complex construct that creates a vicious circle

The identified causes of aporophobia do not act in isolation, rather feed into one another creating a complex construct that is difficult to combat since it has its roots in our biological heritage, has adapted to our social and economic systems, has an impact on our cultural and moral framework and a result in our psychological well-being (Fig 5).

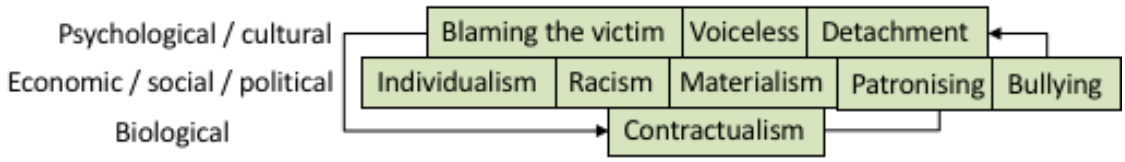

Figure 5. Aporophobia is a complex construct based on biological heritage but built upon the economic and social and has an impact on the psychological and cultural wellbeing. It creates a vicious circle aggravating the stigmatization of the poor. Source: author's creation

As Cortina states (2017), people are biologically programmed for mutual help and, in a market economy, individuals that do not possess wealth are less interesting. However, it is under the capitalist achievement principle and recognition order where aporophobia is aggravated by the element of blame, which exacerbates social stigma and personal shame of the poor .

Identifying the phenomenon of aporophobia and all the circumstances that sustain it is the first step to combat this discriminatory practice. However, the vicious circle created by aporophobia can only be tackled by working on the beliefs and values that are behind the capitalist recognition order and the attribution of blame to the poor.

2.7. Conclusions

Aporophobia is a key topic of our times. In spite of the efforts to reduce poverty in terms of redistributive justice, 700 million people (10% of the global population) still live in extreme poverty and evidence suggests that global poverty could increase in as much as 8% as a result of the COVID-19 crisis, according to the United Nations (UN). While traditional redistribution strategies have proved ineffective in the last decades, this study provides a new path for the analysis of poverty and inequality reduction with focus on the discrimination of the poor, therefore looking into the biological, psychological, cultural, social, political and economic circumstances inherited in our social framework. This approach switches the focus from the poor to the non-poor and the society as a whole. In addition, this chapter proposes topics to be considered for policy making in order to reduce poverty and inequality by mitigating aporophobia, such as the development of new ways of communitarism and intercultural social network, the improvement of the quality of democracy with deeper discussions on poverty, the value of activities such as care and knowledge which are essential, but are not at the center of the existing productivity model, the development of alternative industrial and economic models which are human and environment centered, as well as greater consciousness on minimal global ethics and justice, away from economic lobbies.

Further studies should be carried out to validate and develop this body of knowledge. In particular, the connections between the different causes of aporophobia deserve careful analysis, since they feed one another. On the other hand, additional research is required to move towards the definition of specific policies to combat aporophobia. In this line, studies identifying and quantifying the phenomenon and its implications at a macro-global, meso-national and micro-personal levels would shed some light on the measures to be implemented, starting with public discussion and awareness as first steps to incorporate aporophobia into public and private policy making.

Chapter 3. There are no shortcuts to fairness. An analysis of bias against poverty in AI capitalism

Summary of the chapter

Despite the growing concern about AI systems being biased in critical areas such as healthcare, judicial system, police forces and education, the implementation of ethical guides, standards and regulations have not guaranteed a fair AI. By explaining the nature of bias and fairness, this chapter unveils why AI bias cannot be solved by only dealing with technological development. Focusing on a specific type of bias (against the poor), the paper goes to the root of the contextual reasons that lead to bias and describe the relativeness of the perception of fairness. The paper also explains the process of legitimacy that allow victims of bias to claim for social justice and how AI transforms existing biases by introducing new value recognition orders. Working towards AI fairness would mean higher civil society participation in the power structure of the AI business, highly concentrated on an oligopoly of platforms, and the possibility to take informed decisions as regards the always imperfect ethical trade-offs in terms of fairness, which should be communicated transparently to users.

Key words: Artificial intelligence, capitalism, bias, fairness, poverty

3.1.Introduction

The issue of bias within AI systems has been recognized by policymakers and academics around the world and it is widely documented in the literature, since it reproduces and often amplifies historical types of prejudices (Bolukbasi et al. 2016; Nikhil Garga, Londa Schiebingerb, Dan Jurafskyc 2018; Manzini et al. 2019; Nadeem et al. 2020) with the eventual increase in discrimination (Vinuesa et al. 2020). As a result, regulatory efforts worldwide are being made

towards a responsible development and use of AI (White House 2016; HLEGAI 2019; SCMP Research 2020; European Commission 2021). At the moment, significant attention is being dedicated to the development of assessment tools, such as the assessment framework of the EU Guidelines for trustworthy AI (HLEGAI 2019), created as a result of a public consultation process. It is also being analysed how AI can either contribute or inhibit the achievement of the Sustainable Development Goals (Vinuesa et al. 2020) and it has been suggested that AI has the potential to influence all 17 goals, contributing positively to 134 targets while it can also inhibit 59 targets. At the moment, over 600 AI-related policy recommendations have been released by inter-governmental organisations, professional bodies, national-level communities and other public organisations, non-governmental and private for-profit companies (Dignum 2022). Among this myriad of initiatives, a recent study suggests that there is a global convergence around five ethical principles in the management algorithmic decision making: transparency, justice and fairness, non-maleficence, responsibility and privacy (Jobin et al. 2019).

AI fairness and bias have therefore been widely accepted as a principle in the AI regulatory framework and have also been extensively analysed from the technical point of view of the training data (Rudinger et al. 2018; Chiappa et al. 2020), algorithms (Dwork et al. 2011; Hardt et al. 2016; Kroll et al. 2017a; Green and Hu 2018; Card and Smith 2020; Sap et al. 2020; Jiang et al. 2021), and development practices (Floridi and Taddeo 2016; Floridi 2019a; Vakkuri et al. 2020; Morley et al. 2020, 2021b) , aiming to assist AI developers and designers to translate fairness into an operational reality and consider AI ethics “as a service” (Morley et al. 2021a). However, most existing studies provide micro analyses which focus on technical solutions and have an instrumental approach to fairness, aimed at developing operational AI systems that seek to adapt, mainly, to western societies and values (Carman and Rosman 2021). A multidisciplinary approach of AI fairness is required, since AI not only reflects the inherent bias in the societies where it is trained (West et al. 2019; Vinuesa et al. 2020), in an unsupervised manner (Radford et al. 2019; Talmor et al. 2021), but also has been supporting and exacerbating the existing power

structures of liberal capitalism, colonialism and patriarchy throughout digital revolution (Fuchs 2018; Zuboff 2019; Coeckelbergh 2022). AI can be better understood as part of a "socio-technical ecosystem" (Ausín and Robles Carrillo 2021; Dignum 2022)

Current AI systems do not have the reasoning nor contextualisation capacity; they do not have common sense (Mántaras 2020). However, it is not only by creating increasingly intelligent systems that ethical problems will be solved. The approach to fairness circumscribed within AI can be understood from a transhumanist perspective characterised by technological optimism, where society is thought to change positively as a result of technology, which can therefore play a role to rectify inequalities (Fuchs 2020). However, the approach to fairness from the mere technical perspective only scratches the surface of underlying fundamental social inequalities (Zajko 2021). Blodgett et all (2020) analysed 146 papers studying bias in Natural Language Processing (NLP) systems (published prior to May 2020) and argued that these studies do not provide a conceptualisation of bias outside AI systems. Card & Smith (2020) highlight that literature on fairness within Machine Learning (ML) depends mostly on assumptions. There are no shortcuts to fairness and an increasing number of scholars underline the need for a contextualised analysis for a socially grounded and engaged perspective of AI bias and fairness (Green and Hu 2018; By et al. 2019; Kusner and Loftus 2020; Zajko 2021; Dignum 2022).

This chapter aims, first of all, to discuss the phenomenon of bias, from a psychological and contextual points of view, in society and in AI in particular. Then, it offers a specific analysis of the characteristics that aggravate bias against the poor in the context of liberal capitalism, which is at the root of bias against the poor in AI, and the reasons why not enough studies have been devoted to this type of unfairness. Finally, the paper seeks to explain how AI transforms existing bias against the poor as a result of new barriers and dynamics of AI capitalism. The conclusions propose ways to move forward by sharing responsibility about decisions on fairness and bias with all stakeholders and being able to carry out public discussions and informed decisions on AI

systems as a result of transparency and explainability, not only understood from a technical point of view, but in the sense of facilitating to access the AI narrative (Coeckelbergh 2021).

3.2.What bias and fairness mean

Biases are grounded on our beliefs and are part of human cognition (Allport 1954; Reicher 2007; Pettigrew 2020; Paolini et al. 2021). Since there is not sufficient time to analyse other human beings in all their details and background, we interpret the world that surround us by putting information into categories and generalising from previous experience. Allport describes prejudices as overgeneralised and erroneous beliefs, when these have a social category, they are considered stereotypes and, when they are transmitted through language, they are known as biases, permitting prejudices to be socially shared and to perdure over time (Maass 1999; Beukeboom and Burgers 2019). Therefore, bias is inherent to language and "debiasing" AI is based on the fantasy that there is a neutral value-free environment, when, in fact, it is meant to align with the dominant scientific, social and political values (Green 2020).

Fairness has been dealt with in AI literature mainly from a mathematical perspective, following different approaches. One of them is the so-called "fairness through unawareness", which means making historically discriminated groups "invisible" to AI models. However, this method has been qualified as ineffective due to the existence of redundant encodings (ways of predicting protected attributes from other features) (Hardt et al. 2016; Card and Smith 2020). "Demographic or statistical parity" is another approach to fairness through AI, but it cares about group fairness rather than individual fairness (Dwork et al. 2011). "Individual fairness", argues that models must make similar predictions for similar individuals, has also some drawbacks, since the effects of this framework is highly dependent on the particular notion of similarity that is chosen (Green and Hu 2018). "Randomisation" follows the idea that a policy should not try to protect an attribute, but just ensure that some basic criteria are met (Kroll et al. 2017a). However, all mathematical

approaches are only imperfect approximations and it has been admitted that none of these approaches include all functions that would be morally relevant (Card and Smith 2020).

Coeckelbergh (2022) provides some insights on how the main theories of fairness in social sciences could be translated into AI environments. A common way to explain fairness in welfare states is in terms of a maxmin approach to distributive justice (Rawls 1971), which would require, for example, that algorithms would give priority to individuals, for instance, who live in worse areas, when selecting candidates for a job. An identitarian approach to justice acknowledged by Fraser and Honneth (2003) in an AI context could mean considering positive discrimination of historically discriminated groups by design, as suggested by the "algorithmic reparation" approach (Davis et al. 2021). There is also a meritocratic conception of fairness, critically described by Sandel (2020) and according to which algorithms should need to be capable to track merit. Or under the perspective of the capability approach, fairness would require that everyone should have sufficient basic capabilities (Sen 2001; Nussbaum 2012); therefore, algorithms should find out what people's aims in life are, in order to support them achieving their goals. So, should AI provide positive discrimination of specific groups? Prioritise individuals that have less opportunities? Support individual agency?

In fact, there is not a universal and absolute understanding of bias and fairness. Honneth talks about the "perception of unfairness", which is an experience of social discontent according to expectations within a current institutional order (2003). Therefore, AI should consider fairness and bias within specific contexts and users, bearing in mind that AI systems will always be imperfect and have trade-offs (Whittlestone et al. 2019). As a result of the relative and contextual nature of "perceived fairness", all discriminated groups cannot be analysed as a block, which is often the case within AI literature. Although all types of discrimination have common psychological aspects among the discriminators (who often share a point of view of deservingness of the victims) and the discriminated (who suffer a social stigma (Goffman 1963)), discrimination

in terms of gender, race, nationality or sexual orientation exist as a result of specific power structures and historical contexts. This chapter provides a specific analysis of bias against the poor within the so-called "AI capitalism" (Coeckelbergh 2022) or "surveillance capitalism" (Zuboff 2019), informs about its roots, how liberal capitalism aggravates the phenomenon and the reasons why this type of discrimination has not been sufficiently analysed in literature.

3.3.The nature of bias against the poor

All bias phenomena are contextual, since they appear and develop as a result of historically institutionalised principles of recognition (Fraser and Honneth 2003). Aporophobia, as an expression of prejudices against the poor, acquires a specific connotation of blamefulness with the birth of industrial capitalism, where individuals were no longer socially valued by the codes of honour corresponding to the hierarchies of the feudal regime. Recognition was democratised by according all members of society equal respect for their dignity as a result of individual achievement, as a productive citizen within the industrially organised division of labour (Piketty 2020).

The paradigm of equal opportunity is at the very core of liberal capitalist welfare states and redistributive justice is the main political strategy to tackle poverty (Sandel 2020b). The logic is that individuals are thought to have the possibility to escape their needy situations by climbing up the ladder in line with the concept of meritocracy. Sandel argues that, according to the rhetoric of meritocracy, individuals that work hard are thought to be the ones that deserve and accomplish social and economic success. However, the rhetoric of equal opportunity has also been associated with an overestimation of individual responsibility that leads to the depreciation of the poor, who are considered blameful for not prospering in a context that encourages social mobility (Young 1964; Anderson 1999; Sandel 2020b). Under the rhetoric of meritocracy, the welfare state is no longer a support to individual responsibility but can become a controller of such responsibility (Mounk 2017), where policy makers are forced to justify which poor are victims of bad luck, and

therefore deserving support, and which are not deserving aid ("luck egalitarism") (Anderson 1999).

Unfortunately, this liberal capitalist construct to achieve success and social recognition is more an illusion than a reality. In an era of increasing inequality, the well-known "elephant curve of global inequality 1980 - 2020" documented by Chancel and Piketty (2021), illustrates that the top 1% income distribution captured 23% of total world growth vs 9% for the bottom 50%. Inequality has boosted even more as a result of COVID-19 crisis, according to Gini coefficient estimates. In terms of social mobility, Chetty et all (2014) inform that only 7% of the population of the United States within the 20% lower incomes manage to reach to the 20% top rents within their lifetime. Some European countries, as it is the case of Germany, show even lower social mobility than the US (OECD 2018). But not only the results of the principle of equal opportunity are deceiving. To start with, the principle in itself can be considered unachievable from a conceptual point of view, since every person is inevitably confronted to different experiences even from the moment of birth, so there is no perfect equal opportunity in practice (Fishkin 2014). This shared rhetoric, though, is again a breeding ground for reinforcing the component of blamefulness to bias against the poor, especially in the US where citizens overestimate the potential of social mobility as compared to Europeans (Alesina et al. 2018). Bias against the poor, therefore, would be even more harmful due to the belief of "undeserving" poor (Arneson 1997; Applebaum 2001; Everatt 2009; Nunn and Biressi 2009).

A key contribution of this chapter is the claim that aporophobia is aggravated in the context of "capitalism institutionalized recognition" order based on achievement (Fraser and Honneth 2003 : 151) since it incorporates a sense of blame for being poor (Fig 1). The welfare states rhetoric of equal opportunity, considered the ladder to escape from needy situations (Sandel 2020a), reinforces the sense of shame and self-depreciation on the side of the poor and deservingness on

the side of the non-poor. Cortina states that "equal dignity and compassion are two key elements in the reason of cordial ethics, which are not negotiable in order to overcome this world of inhumane discriminations" (2017 : 27) and we can consider that the psychological consequence of capitalism "individualistic achievement principle" (Fraser and Honneth 2003 : 147) constitutes a kind of anaesthesia for the feeling of compassion towards the poor. At the same time, it generates a feeling of shame that undermines the dignity of the affected individuals.

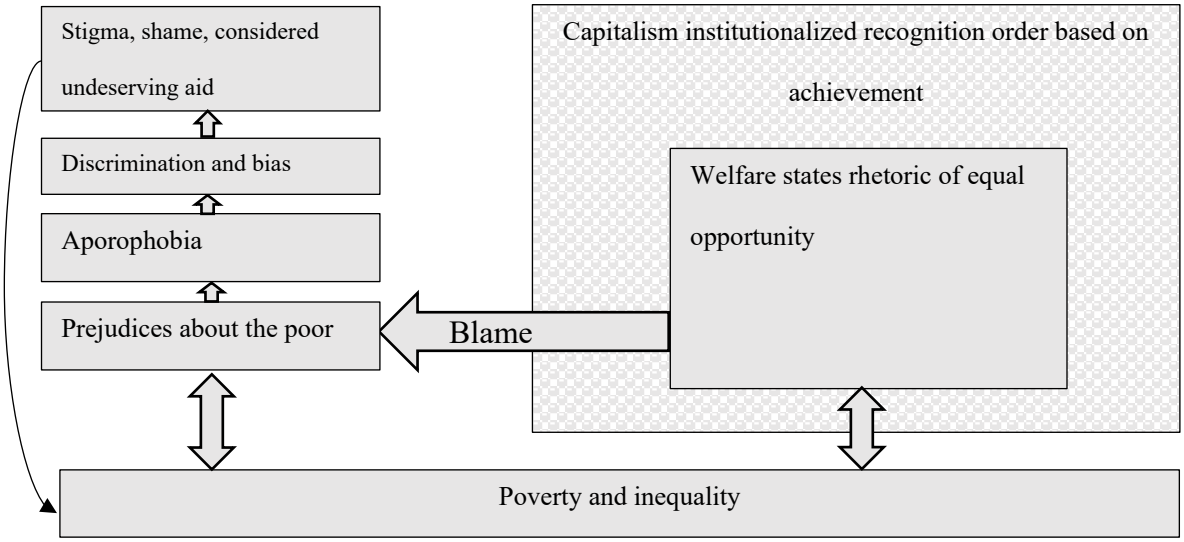

Figure 1: The effects of aporophobia are aggravated under the capitalism institutionalized recognition order based on achievement and meritocracy, since blame is incorporated to the prejudices against the poor, worsening social stigma and transmitted through language as bias.

While in capitalist welfare states, recognition order involves that minimum economic resources have been assured to all individuals, the rest continues to legitimise an extremely unequal distribution of goods according to the capitalist individual achievement principle. Despite individuals having equal legal rights and duties according to the legal framework, social relations are based on the achievement principles and the competition for professional status. However, unlike other historically discriminated groups, aporophobia has hardly received the attention it deserves neither in AI Ethics nor in recent social sciences literature (Cortina 2017).

3.4. Why bias against the poor has not been sufficiently analysed in literature

The topic of bias against the poor can be analysed with from the point of view of fairness as justice in liberal capitalism (Rawls 1971), the perspective of identity (in the so-called identitarian politics), and the class struggle (in Marxism (Fraser and Honneth 2003; D. Grusky 2014)). While bias against the poor does not receive the deserved attention under liberal capitalism, one would think it would be sufficiently studied as part of identitarian politics or Marxist theories. Unfortunately, this is not the case.

First of all, Marxist approaches to class struggles have a universalist approach and do not focus on recognition, but on the underlying economics. From the point of view of Marxian theory, the liberal approach to justice focuses on abstract principles and does not go to the root of unfairness: the actual capitalist structure of society. Bias, therefore, should be abolished by eliminating the structure of capitalism all together and, obviously, Marxist theories do not contemplate recognition of bias against the poor, since there should be no such capitalist social order (D. Grusky 2014; Coeckelbergh 2022). There are, however, more elements related to Marxism that contribute to the current lack of interest in bias against the poor: Fuchs suggests that post-modern theory has an anti-Marxist bias (2020). While in 1968 Marxist theories were being considered, the economic crisis of the 1970s led to neoliberal politics as the world's dominant political paradigm, in a regime that values accumulation, difference, identity and networks (Harvey 2005), as opposed to focus on solidarity, class and the modes of production of Marxism. Therefore, one of the reasons why discrimination against the poor is unaddressed is because the Marxist tradition that endowed the working class with a privilege status of moral discontent in capitalism is nowadays just an unaddressed residue of historical speculation (Fraser and Honneth 2003). Currently, in order to reinvigorate distribution struggles, there should be, at least, what Honneth describes as the "expectations of recognition". However, this is not the case when governments are considered technocratic bodies at the service of market conditions, only aiming to implement redistributional measures negotiated in the form of wages bargaining and tax policies (Sandel

2020b), instead of leading a real debate questing the "capitalist recognition order" (Fraser and Honneth 2003 : 151). .

While in post-modern liberal capitalist societies there is widespread acceptance of bias against the poor, the current debate on bias and fairness focuses on the so-called identitarian thinking (Fukuyama 2018). Since Hegel, the identitarian approach started to shift focus towards historically situated and locally constructed identities (Fraser and Honneth 2003) that claim for recognition to their different value convictions and lifestyles (Taylor 1931). Why are the poor not claiming for that identity recognition? Because the poor would need to feel the legitimacy to do so, which is not the case in the current anti-Marxism biased liberal capitalism recognition order. In addition, it has been argued that this post-modern politics does not challenge capitalism, but coexists with it. As a result, it focuses on identities which are compatible with neoliberal ideology, as opposed to defending the interests of a specific socioeconomic class. In this context, Marxist theories are described as totalizing and reductive, blind to patriarchy, racism and cultural diversity (Dyer-Witheford 2019). According to Fukuyama, societies guided by identity have a difficulty for collective action, since they are fractured into segments (2018). Today focus on multiculturalism (politics of identity) is dominant, where cultural minorities struggle for recognition. Traditional problems of capitalist societies are no longer held to be the key to present moral discontent. In this context, even weak social efforts in terms of "social struggle" are considered a private affair and are "legitimately" excluded from public debate (Fraser and Honneth 2003).

3.5. Transformed biases: what determines value and recognition in AI capitalism

While in industrial capitalism what is distinguished as "work", understood as individual and quantifiable achievement in terms of productivity, becomes a determination of value, in AI capitalism individuals are also valued as a result of being "information hubs". That is the case

because in AI capitalism raw materials are data, AI are the means of production and the new commodities are processed data in the form of individual profiles that allow to attract user's attention, which is sold to advertisers (the new B2B) or to service customers (such as risk assessment companies) through data brokers (the new distributors) (Poell et al. 2019; Zuboff 2019). The persons that act as information nodes are the new achievers in the post-industrial era, since they are the ones that manage to lead towards higher user's attention.

Before going into detail about the impact of AI capitalism on fairness and bias, a brief analysis of the so-called AI capitalism (Coeckelbergh 2022) or surveillance capitalism (Zuboff 2019) is required. Starting with data, it has been suggested that a new social order is settling, which has been called "data colonialism" (Couldry and Mejías 2019; Kwet 2019). Considering that historical colonialism implies the appropriation of resources, the development of unequal social and economic relations, the unequal global distribution of the benefits of resource appropriation and the proliferation of ideologies that justify the appropriation, we can certainly consider we are living under a new kind of colonialism, where data is the new raw material that capitalism aims to control. The resulting ethical question, though, does not only regard the appropriation of data for profit, but the fact that this is done beyond the control of the person to whom the data relates. While historical colonialism expanded through geographical territories, data colonialism gradually conquers different layers of human life, modifying human behaviour, from which new data is collected into the system (Zuboff 2019). This, of course, has long term consequences. While the accumulative characteristic of industrial capitalism has generated a global environmental crisis, one could expect that AI capitalism, which concentrates and deals with knowledge and human experiences, can ultimately lead to a gradual loss of human agency, considering Zuboff's description of how AI can modify human behaviour and go unnoticed (2019).

In terms of means of production, AI capitalism has facilitated the concentration of power into big platforms that share the market of global data commodification (Dijck et al. 2018). In fact, the practical totality of online world traffic is under the control of 8 platform companies, namely Google, Apple, Facebook, Amazon and Microsoft in the western world and Alibaba, Baidu, and Tencent in China (De Kloet et al. 2019). These platforms not only erode the very foundations of traditional capitalism and welfare states as a result of their monopolistic approach and concentration of power globally, but also introduce new and intricated business models of data trade among data platform owners, data brokers, advertising companies, and service companies (Srnicek 2016). Economy has become "platformisated" (Poell et al. 2019).

From the point of view of labour, the first consideration is that direct AI labour, which allows to commodify data and generate personal profiles, is for free. In fact, Fuchs considers that labour can be classified as: wage labour, reproductive labour, slave labour and AI labour (2018) and Qiu describes AI capitalism workers as iSlaves (2017). While individuals are granted free access to apps, details from their health, education, informational consumption and social life are used to generate business. In this context, Marxist theories depicting the antagonism between the capitalist class and the working class, where workers are compelled to produce surplus commodities that the capitalists will sell in order to get profits, can be used to explain AI capitalism, with the difference that raw materials and workforce are obtained for free.

While technological development creates the foundations of new forms of cooperation, under the conditions of liberal capitalism these have also implied new forms of exploitation (Fuchs 2018) and a new recognition order, which increases existing biases and generates new ones. First of all, lacking digital skills or not having access to digital infrastructure is, per se, a new form of exclusion. 35% of people in developing countries and 75% of people in the 48-UN-designated Least Developed Countries (LDCs) do not have access to the Internet, as compared to 13% of individuals in developed countries (International Telecommunications Union (ITU), 2020). In

addition to that, in the US, 16% of the population and 23% of adults internationally can be considered digitally illiterate (Mamedova Emily Pawlowski and Hudson 2018), fact that can act as an aggravator of historical types of discrimination (Adeleke et al. 2021). While AI allows for more effective production, it also fosters uneven development of countries and the creation of dual local economies as a result of the so-called digital divide (van Dijk 2020). The psychological consequences of AI capitalism are also distributed unequally among the "connected" individuals. While there is a shared anxiety as a result of the internalised need to achieve (Moore 2019), low-status workers are more highly exposed to surveillance than high-status workers, although their data is also exploited (Couldry and Mejías 2019). It has also been documented that AI capitalism allows for degrading working conditions of the most vulnerable (Azmanova 2020) in what Standing called the "precariat" (2011). In fact, organizations increasingly use customers to directly monitor workers (Maffie 2020) since AI allows to collect customers feedback to evaluate worker's performance in real time. These practices have given some workers a digital "boss" (Vallas and Schor 2020). This is especially true in platform companies, where customers are embedded as a layer to algorithmically control the performance of "platform-based gig workers" (Stark and Pais 2020; Lei 2021; Rahman and Valentine 2021; Cameron and Rahman 2021). One could consider that this is an AI version of "the customer is always right", where customers have complete control over workers, but no accountability for their actions. This, of course, has consequences on the workers incentives, continuity of employment (Maffie 2020) and even future work opportunities (Kellogg et al.). COVID-19 crisis has accelerated things as regards the automation of managerial functions, which goes from hiring to firing, precipitating the precariat (an increase of 26% of automated monitoring of workers has been documented within the month of July 2021 in the US (Kelly-Lyth and Stevens 2022). All in all, AI has been considered to foster the culmination of the alienation of workers under capitalism (Dyer-Witheford 2019).

The enhancement of bias against the poor, however, is even more intricated within the core of AI systems at the service of capitalism. According to Eubanks description of the "digital

poorhouses", the poor are often excluded from health insurances as a result of automated eligibility and can even be categorised as problematic parents (2018). Benjamin's explanation of a "scored society" (2019) describes how the poor are targeted purposely by specific predator financial services while they can be tagged as risky investments (Fig 2). Therefore, one could consider that some AI systems are "biased by design" against the poor, as a result of which needy individuals and families are spotted purposely and excluded from basic services as a result of predictive risk models. When the postmodernist ideology of accumulation, globalisation, stressing identity described by Harvey (2005) has strengthened its network and surveillance capacity, it has also become more burdensome against the poor, who are either expelled directly from the system by not having access to the network or biased in more efficient and opaque ways. An important difficulty to fight against "bias-by-design" is that users not only do not easily realise they are being biased, but they cannot prove it either (Kelly-Lyth and Stevens 2022).

Finding solutions to mitigate bias against the poor in AI, therefore, is not an easy task, since the values sustaining this bias are not dependent on AI networks and information flows, but are rooted in society as a whole in a context and as a result of historical reasons. The problem, therefore, is not AI, but AI capitalism. Solutions to this type of bias, then, should not be found technically, but within economic management, promoting a change of the traditional capitalist hierarchical system now exacerbated by the power concentration of the big AI platforms, assisting governments and civil society to recover the required parcel of decision making that has been lost within the intertwined circuits of data trade and commodification within the gig economy (Dijck et al. 2018). It has been suggested, for example, that AI companies organised as cooperatives run by users would trigger changes in the AI recognition order (Coeckelbergh 2022).

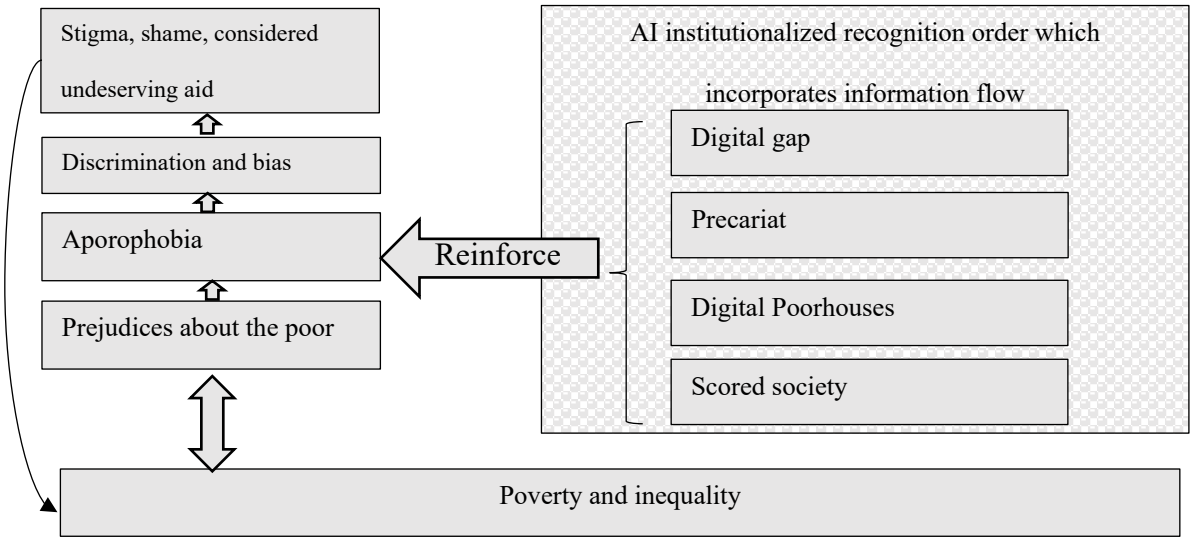

Figure 2: The information flow incorporated in AI institutionalised recognition order reinforces aporophobia as a result of the digital gap (van Dijk 2020), the "precariat" (Standing 2011), the "digital poorhouses" (Eubanks 2018) and the "scored society" (Benjamin 2019).

Having said that, the internationally agreed Trustworthy AI principles, which will be legally enforced through the EU AI Act (European Commission 2021), are a step forward to claim big platforms for explainability and transparency, which would certainly contribute to identify biased practices and take public action to correct them. Societal problems cannot be solved by outsourcing them to big corporations (Dijck et al. 2018), however, big corporations have to be requested for transparency so that public debate can take place on the shared values that AI platforms are supporting. This, nevertheless, poses an additional challenge since AI platforms are operating globally but there is not a "global society" or "global shared values". AI systems mirror western culture neglecting the identity of the Global South. In this regard, ethical pluralism, which acknowledges the presumption of universally valid norms and that different views can emerge from diverse interpretations or applications of shared norms (Ess 2020), may be a way

forward, considering international ethical principles and implementing them to specific culture and context requirements.

3.6. Conclusions

Despite the proliferation of ethics guides (Jobin et al. 2019a), standards (Chatila and Havens 2019) and auditing procedures (Mökander and Floridi 2021; Ugwudike 2021), the fast growing business of anti-bias training (Eberhardt 2020) and the regulatory efforts being put in place (White House 2016; SCMP Research 2020; European Commission 2021), ethically questionable AI systems continue to grow in areas such as healthcare (Raghu et al. 2019), judicial system (Kleinberg et al. 2018), police forces (Ratnaparkhi et al. 2021) and education (Newton 2021). It has been suggested that one of the reasons why there is no real engagement towards AI ethics from AI development teams and big platforms is because of the lack of incentives (Hacker 2018). But even if AI developers wish to engage personally with the topic, it is difficult to do so if what you are hearing about AI ethics is unclear, accusatory and threatening (Eberhardt 2020). The perceived lack of AI fairness is decreasing the perceived legitimacy in the algorithmic decision making (Martin and Waldman 2022) and generating a lack of trust in AI (von Eschenbach 2021). This is a particularly serious matter when the Edelman Trust Barometer (2020) reports that trust in technology has declined in 21 of 26 markets surveyed and less than 50% of respondents in the US, Canada, the UK, Germany, France and Ireland reported trust in AI. In fact, only 44% of survey respondents globally believe that the use of AI will have a positive impact. This, of course, can constitute an obstacle for the successful development and implementation of AI, which is a very valuable tool to work towards some of the biggest global challenges (Vinuesa et al. 2020).

Clearly, the responsibility given to AI development teams might be overwhelming, when, in fact, there is no consensus on the meaning of fairness, which is dependent on the historical context and individual perception. Therefore, the expectation to find completely unprejudiced AI systems

means either a naïve approach to technology, expected to automatically bring positive social progress, or the belief that western liberal capitalist values are shared worldwide and in the same exact manner for each individual. In the meantime, while AI paradigms reinforce existing power structures, the main analysis focuses on design and operationalisation of AI systems, seeking to translate abstract ethical principles into practices (Floridi 2019b; Ibáñez and Olmeda 2021; Morley et al. 2021c) but there is not an analysis of how AI actually affects society (Dignum 2022).

There are, however, ways forward to tackle the relevant topic of AI bias and fairness. First of all, AI systems have the potential to provide us with a useful mirror where to identify bias from big amounts of spontaneous data found in word embeddings pretrained from social networks and news (Joseph and Morgan 2020). Therefore, AI can constitute a useful tool to measure and follow up different types of bias in line with the contextual world events, helping us analyse the historical context of bias in real time. Secondly, the societal, psychological and economic changes generated by AI platforms and new business models need to be studied in order to identify new forms of exclusion that can add up to the burden carried by historically discriminated groups, either in terms of identity or socio-economic factors. One could imagine a future where there is a diversification of AI platforms, some of which could be civil-society-based cooperatives, that openly communicate the ethical decisions taken in terms of fairness and bias according to transparently publicised ethical principles and cater for a diversity of value-oriented users who, based on information provided by AI platforms, take informed decisions about what platforms decide to use. One could also imagine specific ethically tagged content within big platforms, to cater for a diversity of value-oriented users (within a global set of agreed values) according to political view and local traditions, catering also for the Global South.

Finally, none of the above can be implemented without understanding the always imperfect trade-offs that AI development teams need to make when dealing with the topic of bias and fairness. Therefore, the technical perspective of AI bias and fairness is crucial in terms of providing

transparency and explainability, since accountability for fairness and bias cannot be shared with all social stakeholders that participate within the AI value chain if AI is designed in the form of opaque systems and, most importantly, this is not openly communicated. In this sense, the legal enforcement of a "strong explainability requirement" for organisations that decide to automate decision making, suggested by Maclure (2021), seems to be a possible step towards whenever is technically feasible, at least, be able to evaluate fairness. The now incipient emergence of data management intermediaries, which seek to fill in the trust gap by managing data throughout the AI value chain according to users' requirements (World Economic Forum 2022), is a concrete example of the added value that transparency and explainability can provide. In other words, it is time for the black box to open (as far as it is technically possible) to a more participative algorithm decision-making process, which can constitute an opportunity for new business models and stakeholders' roles that focus on increasing digital agency.

Chapter 4. Are AI systems biased against the poor? A machine learning analysis using Word2Vec and GloVe embeddings

Summary of the chapter

Among the myriad of technical approaches and abstract guidelines proposed for the topic of AI bias, there has been an urgent call to translate the principle of fairness into the operational AI reality with the involvement of social sciences specialists in order to analyse the context of specific types of bias, since there is not a generalizable solution. This chapter offers an interdisciplinary contribution to the topic of AI and societal bias, in particular against the poor, providing a conceptual framework of the issue and a tailor-made model from which meaningful data is obtained by using Natural Language Processing (NLP) word vectors in pretrained Google Word2Vec, Twitter and Wikipedia GloVe word embeddings. The results of the study offer the first set of data that evidences the existence of bias against the poor and suggest that Google Word2vec shows a higher degree of bias when the terms are related to beliefs, whereas bias is higher in Twitter GloVe when the terms express behaviour. This article contributes to the body of work on bias, both from and AI and a social sciences perspective, by providing evidence of a transversal aggravating factor for historical types of discrimination. The evidence of bias against the poor also has important consequences in terms of human development, since it often leads to discrimination, which constitutes an obstacle for the effectiveness of poverty reduction policies.

Keywords: Bias, Artificial Intelligence, embeddings, poverty.

4.1. Introduction

It is widely documented that Artificial Intelligence (AI) reproduces and often amplifies biases against historically disempowered groups (Bolukbasi et al. 2016; Nikhil Garga, Londa

Schiebingerb, Dan Jurafskyc 2018; Manzini et al. 2019; Nadeem et al. 2020). This constitutes a risk for the exacerbation of those biases offline and the eventual increase in discrimination (Vinuesa et al. 2020). AI systems are not ethically neutral but, more and more, we are all dependent on AI for our decisions (Fry 2018). In the information society, AI is at the core of high risk services such as healthcare (Watson et al. 2019; Zetterholm et al. 2021; Vallès-Peris and Domènech 2021), financial services (Kostka 2019; Townson 2020; Lee and Floridi 2020; Aggarwal 2020; Anshari et al. 2021) justice and security (Poitras 2014; Hauge et al. 2016; Merler et al. 2019; Green and Chen 2019) and even the military (Vynck 2021). AI is also an integral part of marketing, predicting users' interests through big data that contains each person's personal digital profile, in what has been called "surveillance capitalism" (Zuboff 2019).

While the amount of algorithmic systems performing in questionable ethical manner continues to grow (Tsamados et al. 2021a), governmental efforts to regulate AI have gained momentum (White House 2016; SCMP Research 2020; European Commission 2021). At a regional level, the European Union is considered to have an ethically superior regulatory framework in terms of citizens' rights (Allison and Schmidt 2020; Gill 2020; Imbrie et al. 2020; Roberts et al. 2021), which has a positive impact at a global level (Bradford 2020). At the core of the EU AI framework, there is the principle of "diversity, non-discrimination and fairness", including the "avoidance of unfair bias", especially in the case of the historically discriminated groups (HLEGAI 2019). However, the legal framework is not sufficient, considering that the ethical principles contained in the law are described as too abstract to implement in practice, often leading to some counterproductive practices such as ethics shopping, ethics bluewashing, ethics lobbying, ethics dumping or ethics shirking (Floridi 2019a).

There is a growing agreement on the urgent need to know how to translate this general ethical framework into the operational AI development (Floridi 2019b; Vakkuri et al. 2020; Morley et al. 2021a, b). In this context of "moral panic" (Ess 2020), there has been a proliferation of AI

ethics guidelines (more than 173 documents in existence in 2021 (Algorithm 2021)), there is a panoply of strategy proposals to detect and correct bias in the data of AI NLP systems (Bolukbasi et al. 2016; Nikhil Garga, Londa Schiebingerb, Dan Jurafskyc 2018; Manzini et al. 2019; Nadeem et al. 2020; Zhao et al. 2021), incipient attempts to train algorithms to detect bias (Sap et al. 2020; Jiang et al. 2021) and algorithmic mathematical constructs which try to achieve partial approximations to fairness (Dwork et al. 2011; Hardt et al. 2016; Kroll et al. 2017a; Green and Hu 2018; Card and Smith 2020).

However, in order to translate the principle of AI fairness (HLEGAI 2019; European Commission 2021), into an operational reality, an in-depth analysis is required, far from the existing turmoil of quick-fix solutions. Bias within AI systems is only the tip of the iceberg, since AI reproduces the prejudices of the societies where they are trained (West et al. 2019; Vinuesa et al. 2020) in an unsupervised manner (Radford et al. 2019; Talmor et al. 2021), either within the data (Rudinger et al. 2018; Chiappa et al. 2020), the algorithms (Mittelstadt et al.; Tsamados et al. 2021b) or even as a result of development procedures (Floridi 2019a; Vakkuri et al. 2020). Therefore, trying to solve the AI ethical problems only through a technical approach is clearly insufficient, since it has only a superficial impact on fundamental inequalities (Zajko 2021). Blodgett et all (2020) analysed 146 papers studying bias in NLP systems (published prior to May 2020) and concluded that these papers do not provide an actual conceptualisation of bias outside NLP systems. Card & Smith (2020) suggest that literature on fairness within ML depends mostly on assumptions. A growing number of voices highlight the need for involvement from the social sciences perspective (Green and Hu 2018; By et al. 2019; Kusner and Loftus 2020; Zajko 2021) since bias needs to be discussed in the “onlife”, using Floridi’s term 2015). In fact, the aim to debias AI systems is based on the illusion that there is a neutral value-free environment, when it is really meant to align with the dominant scientific, social and political values (Green 2020).

When we analyse the nature of bias, it becomes evident that we cannot draw a hard line between what is sufficient and insufficient proof of it, since it is based on our beliefs and a characteristic of human cognition (Allport 1954; Reicher 2007; Pettigrew 2020; Paolini et al. 2021). In fact, the reason why human beings are not only perceived based on their individual characteristics is because we do not have enough time to understand every single detail of every person. Therefore, we put information into categories and generalise based on previous experience. Overgeneralised and erroneous beliefs lead to prejudices. When prejudices have a social category, they are described as stereotypes and, when they are transmitted through the linguistic process, we know them as bias, generating a self-perpetuating cycle in which prejudices are socially shared and maintained (Maass 1999; Beukeboom and Burgers 2019). While bias is the linguistic expression of shared social prejudices within a specific culture, discrimination has been defined as an action of exclusion as a result of prejudice (Allport 1954).

But seeing the tip of the iceberg (bias in AI systems), also tells us that there is an iceberg. Bias in AI acts as a mirror, showing the prejudices that go unnoticed off-line and helping us to evidence an unnoticed discriminatory phenomenon (Hoffmann 2019). While algorithms reproduce inherent tensions at a technical level (Hacker 2018), this data can be used as a warning towards a stigma, which can then be studied from a social sciences perspective since it has a history behind (Zajko 2021). This is precisely what this paper offers: the evidence of bias against the poor in social networks, a neglected type of discrimination in both AI bias and social sciences literature, named "aporophobia" by the philosopher Adela Cortina (2017).

The bias against the poor, which often leads to discriminatory behaviour, has dramatic repercussions since it hinders the effective implementation of poverty reduction policies (Arneson 1997; Applebaum 2001; Everatt 2009; Nunn and Biressi 2009), hampering the work towards the first Sustainable Development Goal of the United Nations (no poverty). It also has a clear impact on the historically discriminated groups (Alessina and Glaeser 2013) and it is closely related to

gender discrimination in capitalist development (Folbre 2021). Sadly, it has been underestimated as a transversal type of discrimination, since there is the tendency within the antidiscrimination discourse towards a single-axis thinking Crenshaw's (1991). However, stereotypes exist within a network of beliefs (Freeman and Ambady 2011), where there is a dynamic interaction among them (Ridgeway and Smith-Lovin 1999) and an aggravating effect for what Hoffman defines as the "multi-oppressed" (2019). Eubanks (Eubanks 2018) identifies algorithms that discriminate the poor and O'Neal (2016) describes how some predatory AI systems target people in need. However, there is no empirical evidence about bias against the poor in the existing literature. This study aims to fill in that gap by offering the first dataset that identifies and measures bias against the poor in the publicly-available Google News Word2, Wikipedia GloVe and Twitter GloVe pretrained word embeddings, providing a study at scale and in context (Joseph and Morgan 2020).

This article offers an interdisciplinary contribution to the topic of AI and societal bias, with special focus on bias against the poor, and it is organised in 5 parts. First of all, it provides an analysis on the roots of discrimination against the poor. Then, we present the materials and methods being used, such as the rationale behind the target terms and attributes that are being searched, the pretrained word embeddings that have been analysed and the methodology to identify and measure bias against the poor using Natural Language Processing (NLP). The key results are then analysed in order to discuss the main implications and conclude.

4.2. The aggravation of bias against the poor under the rhetoric of meritocracy

Redistributive justice is at the very foundation of welfare states, where the principle of equal opportunity and meritocracy are considered to be the main political answer to reduce poverty and an attempt to promote social mobility. But the rhetoric of meritocracy has also been associated with the blamefulness of the poor, who are considered responsible for not climbing up the social ladder (Young 1964; Anderson 1999; Sandel 2020b). The disempowerment and resentment of

the poor is aggravated by the increasing inequality, in particular in the US since the 1980's (Piketty et al. 2018), which has boosted as a result of the COVID-19 crisis, according to Gini coefficient estimates.

The bias against the poor, therefore, is aggravated by the blamefulness associated to this condition in the context of capitalist welfare states and leads to discrimination which contributes to a self-fulfilling prophecy of failure to climb up the ladder (Honneth 1996). Nevertheless, bias against the poor reflects a morally narrow view of social merit, limited to economic and professional credentialism. It is only when the focus is on salary and consumption that badly paid jobs lack social recognition. During the COVID-19 crisis, precariously paid workers in sectors such as delivery and hospital staff enjoyed an increased social recognition, which is essential to overcome the feelings of shame among the stigmatised and the beliefs of deservingness on the side of the stigmatisers (Goffman 1963; Hegel [1807] 1991; Honneth 1996).

By offering evidence about the bias against the poor, this study only scratches the surface of a global and transversal type of social exclusion that potentially can affect 700 M people (10% of the total world population) that currently live in extreme poverty, according to the United Nations (evidence suggests that global poverty could increase by 8% as a result of COVID-19) and is not limited to developing countries (in 2019, 92,4 M people in the EU-27 are at risk of poverty or social exclusion (21.1% of EU-27 population) according to Eurostat).

4.3.Detection of bias against the poor: materials and methods

4.3.1. Materials

A). Target terms and attributes

Bias cannot be treated as a generalizable manner, but in a context (Zajko 2021), for which a framework is required, from the social sciences perspective, to obtain and analyse meaningful data that can be offered by AI. With that purpose, this chapter offers a model to identify and

interpret bias based on Cortina's work on aporophobia (rejection towards the poor) (2017) and Allport's categorization of the degrees of "negative action" associated with prejudices (1954). Cortina uses a list of 17 expressions associated with rejection towards the poor. In our study, we have used 263 synonyms, antonyms and related terms to Cortina's expressions in order to understand how these are related to the concepts of "rich" and "poor". We investigate whether or not a set of favourable attributes are closer or not to the target term "rich" (positive bias towards the rich) and a set of unfavourable attributes more closely related or not to the target term "poor" (bias against the poor). Following Allport's categorization of "negative action" resulting from prejudices, the terms that are part of the study can be grouped into 1 category expressing "belief" (28 favourable and 23 unfavourable words) and 5 categories expressing different degrees of favourable (93 words) or unfavourable attitudes (119 words). The different categories defined by Allport are not sealed compartments, but a conceptual way to organize the favourable and unfavourable expressions that are part of the study and can potentially express bias against or in favour of the poor and the rich.

B). Word coding/Embeddings

We have measured the semantic distance between the 263 favourable and unfavourable attributes related Cortina's expressions and the key terms "rich" and "poor" using vector word representations, which is the state-of-the-art technique in Natural Language Processing (NLP). More specifically, we have observed the semantic relationships between the vector word representations in word embeddings (key terms and attributes) in a simple and intuitive way by using the cosine distance. In our model, we have proposed the use of three types of categories of words, which we have called favourable, neutral and unfavourable attributes, in order to measure the semantic distance to the key terms "rich" and "poor" and measure bias.

The concept of embedding was born as dense vector representations of words or sentences, with the ability to map, syntactic and semantic relations in a vector space, which is core to Natural

Language Processing (NLP) application (Almeida and Xexéo 2019; Camacho-Collados and Pilehvar 2020). Word embeddings are classically classified into two types: count-based embeddings, whose representation is derived from word counts and word frequencies, and predict-based embeddings, which are derived from word context (words neighbouring a core word). The latter are the base of cutting-edge Neural Language Models approach (Adamuthe 2020) and are the most widely used (Gutiérrez and Keith 2019). For our work, we have used Word2Vec (Mikolov et al. 2013a), FastText (Bojanowski et al. 2016) and Glove (Pennington et al. 2014) which are unsupervised approaches based on the hypothesis that words whose occurrence arises in the same contexts tend to have similar meanings. By using this approach, we are able to measure the distance between words/vectors within a context, since the embedding contains the context information of the data used to build it.

The technique we present in this paper could be compared, in a certain way, with a text mining analysis based on an exploratory study where word counting and word clouds could be proposed for a semantic analysis, where the word with the highest frequency is considered to be the most relevant. However, for a study involving millions of different grammars, the task would become very complex to reach relevant conclusions in terms of identifying bias. Besides, we have selected to perform a vectorial study of the numerical representations of the embedding context, because it offers better explainability, required for all approaches based on machine learning models.

C). Pretrained embeddings

We have detected and measured bias against the poor in pretrained word embeddings, which are trained on large datasets and constitute an appropriate and available option to measure the distance between the target terms and attributes of the study. In future studies we aim at training our own embedding, which will allow us to ensure the quality of the data involved and have more control on the amount of context being compared, providing the possibility, for example, to look for bias against the poor not only by using term associations, but also sentence associations. In this study,

we have obtained results from three different embeddings (Google News Word2, Wikipedia GloVe and Twitter GloVe). We have then compared the results, reaching conclusions about the common trends among the three datasets as regards bias against the poor and the specificities of this phenomenon in each embedding.

-Google News word2vec embedding is a pre-trained model of word representation as vectors, using 300 features or coordinates in a 300-dimensional system. This model was trained using a Google News database (about 100 million words). A representation of more than 3 million words and phrases was obtained. The base algorithm used for the creation of this embedding was proposed by Mikolov et al. (Mikolov et al. 2013b). The resulting model has a weight of 1.3 Gb.

-Wikipedia GloVe embedding is a pre-trained word representation model that was trained using the GloVe technique based on the global co-occurrence matrix between words, using as training corpus a dataset of Wikipedia publications. The Wikipedia corpus contains about 2000 million words from 4400 million Wikipedia pages consolidated up to 2014. Additionally, it contains the Gigaword 5 dataset, a comprehensive collection of news text data that has been acquired over several years by the Linguistic Data Consortium (LDC) and contains 4 billion words. The resulting word representation model contains 6 billion tokens, 400 thousand words and was trained with all words uncased. There are four versions of trained embeddings with 50, 100, 200 and 300 vector dimensions. The weight of the resulting model is 822 MB.

-Twitter GloVe embedding is a pre-trained word representation model that was trained using the GloVe technique based on the global co-occurrence matrix between words, using as training corpus a dataset of tweets extracted from the social network Twitter. For the construction of the model, 2 billion tweets written in English were taken. The resulting model contains 27 billion tokens, 1.2 million vocabulary words, and was trained with all words uncased. For this word representation model, there are 25, 50, 100 and 200 dimensional versions. The weight of the resulting model is 1.42GB.

4.3.2. Methods

The following diagram (Fig 1) illustrates the proposed solution to detect and measure bias against the poor by using the key terms “rich” and “poor”, 263 “favourable” and “unfavourable” attributes and vector word representations to measure semantic proximity using the cosine distance in pretrained word embeddings (Google News Word2Vec, Wikipedia GloVe and Twitter GloVe). We have also tested the model using “neutral” attributes. We are fully aware of the limitations attached to the use of some of these attributes, in particular those that work both as nouns and adjectives to other secondary attributes established in different contexts. For this reason, a rich array of expressions was chosen.

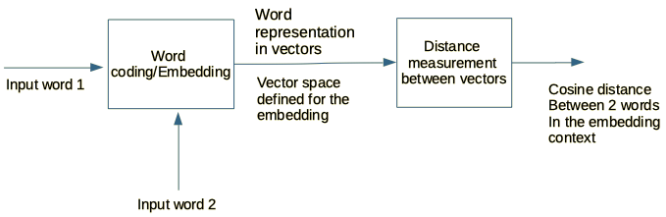

Fig 1 Block diagram of the proposed solution. Source: authors’ creation

- *Semantic analysis of words based on vector distances*

The basis of this work is the semantic analysis based on distance. In order to get reliable information of the relationship between words, we have decided to use the cosine distance, since this numeric metric preserves the relative direction of two vectors inside the vectorial space (in our case, the semantic similarity between words).

- *Cosine distance between words*

The cosine of angle indicates directly proportional similarity between two-word vectors. As the metric increases, it indicates that there is greater similarity between the words. Mathematically, similarity between vectors is defined as the cosine of the angle between the vectors, so the closer the vectors are form an angle to zero, the more similar they are. The cosine of the angle is defined

with the equation (1). Thus, the cosine of the angle is defined as the dot product (indicating the similarity between the two vectors offering a single result) divided by the multiplication of its norms.

$$\cos(\theta) = \frac{A^T B}{|A| \cdot |B|} \quad (1)$$

- *Calculation of the dot product between words*

The similarity metric based on the dot product between the word vectors is directly proportional to the scalar value resulting from the operation. However, this metric increases not only by the cosine of the angle of the vectors, but also by the length of the vectors, so it is necessary to take into account that the metric may be biased by the length of the word vectors. The dot product is defined as in the equation (2):

$$a_1 b_1 + a_2 b_2 + \ldots + a_n b_n = |A||B| cos(\theta) \quad (2)$$

- *Semantic relations between target and attribute words based on cosine distance*

263 registers were built to capture the semantic relationships between the two target terms "rich" and "poor" literally, and the attribute words used as reference points to measure the semantic similarity. It should be taken into account that the value obtained is a number between -1 and 1, since the cosine of an angle belongs to this interval. To carry out our study we have applied the function arc cosine, presented in equation (3), to find the original value of the angle in its natural magnitude radians.

$$\theta = \text{arccosine(similarity cosine)} \quad (3)$$

- *Identifying logical relationships (Analogies) in the same context (embedding)*

A word embedding model can be evaluated on the basis of performance in solving analogy questions. This task was first introduced by Mikolov et al (Mikolov et al. 2013a) and consists of performing additive operations between word vectors. The following equation summarises the so-called "analogy relation" that exists between vector operations.

$$\widehat{\text{rich}} - \widehat{\text{word1}} = \widehat{\text{poor}} - \widehat{\text{word2}} \quad (4)$$

Based on the above, one can seek to predict the vector of one of the words by clearing the equation as follows:

$$\widehat{word2} = \widehat{poor} + \widehat{word1} - \widehat{rich} \quad (5)$$

The result of this equation would be the vector of the word2. In practice, cosine similarity is used to determine that the closest word vector corresponds to the correct answer of the analogy. As a result, we can provide evidence whether a word embedding model is able to maintain the semantic and syntactic relationship between words.

4.4. Results and discussion

The proximity was calculated between the different attributes and the target terms "poor" and "rich". In Table 1 (in the appendix), the relative value of 1 indicates that the attribute is closer to the poor than to the rich in terms of cosine. Alternatively, relative distances can be calculated in radians and then results need to be read the other way round, namely, the longer the distance, the weaker the association between the attributes and the categories of rich and poor.

The main advantage of using radians is that we can calculate "distances of distances" (DD), evaluating the difference between how a certain attribute is associated to the poor as compared to rich, allowing a quantitative expression of the bias net effect, which we have named "aporophobia bias indicator" (ABI). The ABI, therefore, constitutes an intrinsic way to evaluate bias against the poor in pretrained models for given attributes. We have named this model AWEAT (Aporophobia Word Embedding Association Test), since it is inspired on the WEAT (Word Embedding Association Test) by Caliskan et al (Caliskan et al. 2017). The AWEAT allows to order and classify the different attributes from higher to lower ABI for a given pretrained embedding (Google News Word2Vec, Wikipedia GloVe and Twitter GloVe) and find out which negative attributes imply higher bias, since they are more closely related to the term "poor" as opposed to the term "rich". If we consider that the lowest negative ABIs are around 0,14 and that the highest are around 0,5, we can split this interval into quartiles (following the standards of the Human Development Index). The cut-off points are less than 0,02 for low bias, 0,18 for medium bias, from 0,18 to 0,34 for high bias and above 0,34 to very high bias against the poor. This

classification is based on the current selection of attributes. Should the attributes change, the classification should change accordingly.

This order and classification bring meaningful information to the research, since attributes such as “antipathy”, “hate speech” and “hate act” would be classified as low bias (in the sense of the level of association of these attributes to “poor” as compared to “rich” in Google News Word2vec pretrained embedding) whereas, at the other extreme, attributes such as “mediocre”, “dreadful” and “substandard” would be classified as very high bias. Therefore, we should distinguish here between association (distance) and gravity (seriousness) of a construct. In this analysis we are not handling any evidence about the gravity of these attributes. Instead, our focus is on their degree of association (distance) with the poor in the characterisation of bias. For instance, as much as “substandard” seems to present the highest association with the poor, as showed in Table 1, it seems to be a relatively inconsequential attribute if compared to “hate acts” or “insults” in terms of their gravity. It is also interesting to analyse some of the attributes that were originally used by Cortina (2017) to see how they compare to each other in terms of ABI. It is possible to see from Figure 2 that some attributes such as ‘disgust’, ‘disregard’ and ‘fear’ appear to be more closely associated to the poor (meaning that there is a lower relative distance of that attribute in relation to the poor than in relation to the rich) than others such as ‘antipathy’ and ‘aversion’.

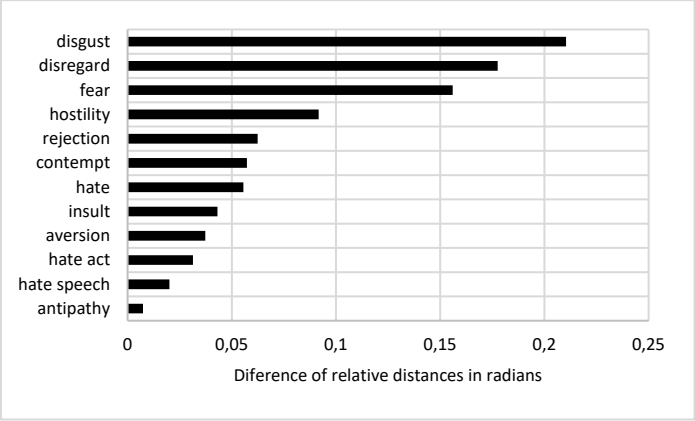

Figure 2: ABIs for unfavourable attributes used by Cortina (Cortina 2017) in Google News Word2vec embeddings. Source: authors' creation.

Our study, however, includes a wider range of negative expressions (other than those mentioned by Cortina) and this unveils a more complex reality. Figure 3 illustrates in grey the attributes used by Cortina and in black a sample of other attributes included in the study, following Allport's categorization of prejudices according to the degree of associated action (Table 2 in the appendix). As a result of broadening the semantic scope and the number of attributes, we find out that attributes that can be included under the categories of "beliefs" or "communication" such as "substandard", "mediocre" or "indifference", according to Allport's categorisation (1954), have clearly higher ABIs (Table 2). In contrast, attributes that have a stronger degree of action, such as "insult", "hate speech" or "hate act", which are associated to Allport's categories of "discrimination" and "physical attack", are more equidistant to the key terms "rich" and "poor" and therefore less closely associated to the poor.

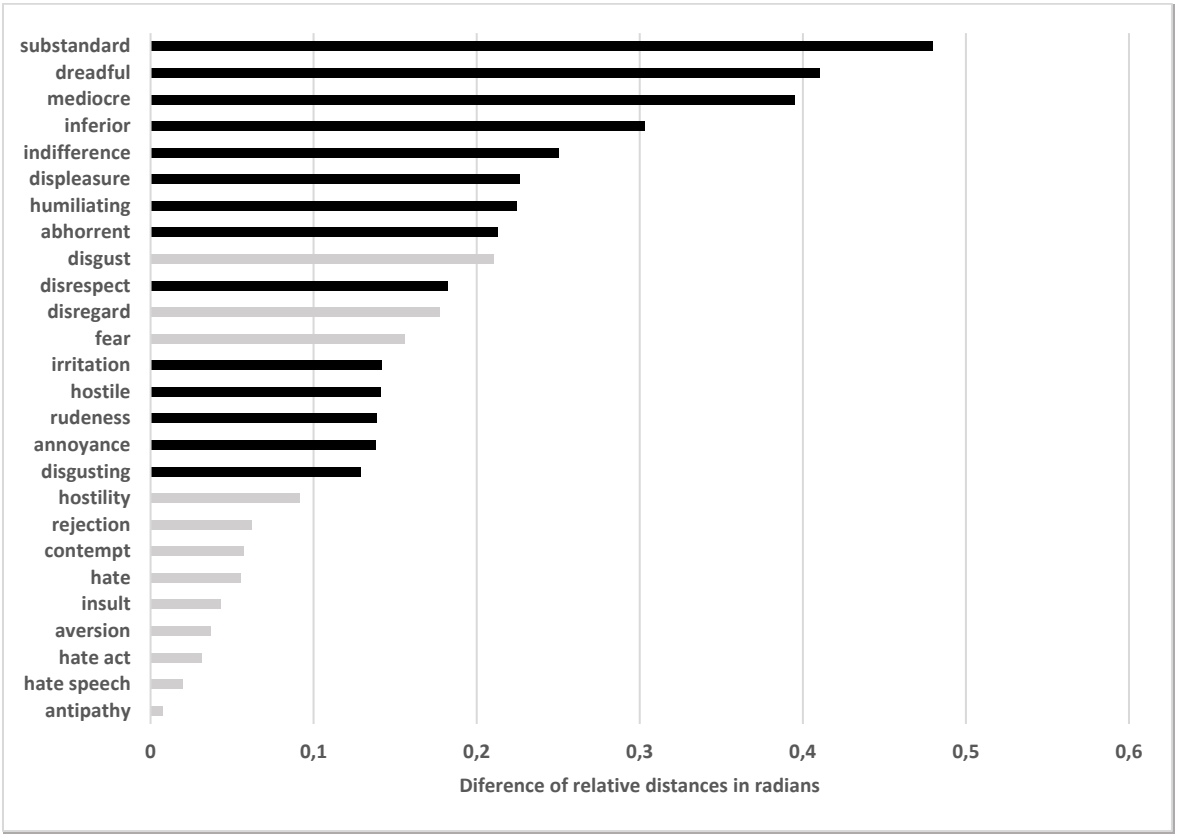

Figure 3: ABIs for unfavourable attributes in Google News Word2vec embedding. Unfavourable attributes used by Cortina (Cortina 2017) are shown in grey. Source: authors' creation.

When analysing the results of the favourable attributes (Table 3 in the appendix), two features are immediately evident from a first inspection. First, results for favourable attributes are not necessarily symmetric to unfavourable attributes (as expected, since the terms themselves are not completely symmetric). Second, some favourable attributes are more closely related to the poor than to the rich, characterising elements that prima facie could be understood as positive bias towards the poor. However, a close inspection reveals that attributes of "sympathy", "politeness", "pleasing", "goodwill", "cordiality" and "friendliness" are all compatible with a certain sense of subservience that can be expected from the poor, reinforcing a certain stereotype of inferiority. We can also verify that some words are relatively neutral towards the rich and the poor. On the other hand, the closer distances found out between favourable attributes and the "rich" reveal hedonist attributes related to attractiveness, pleasure, taste, etc, all part of elements of

‘distinction’, as famously portrayed by Bourdieu (Bourdieu 2010). This phenomenon could be an evidence of plutofilia or overestimation of the rich, which, according to Allport is a previous step to aporophobia, since “one must first overestimate the things one love before one can underestimate their contraries” [8: 25].

It is important to remark, however, that Google News Word2vec pretrained embedding is not the only informational basis that has been used for this assessment. Two additional embeddings, trained on different databases, are integral part of the study, namely Twitter Glove and Wikipedia Glove. The coincidences between the three analysed embeddings provide robustness to the AWEAT model. Figures 4, 5 and 6 display the key results.

In Figure 4, positive results indicate that the ABI in Google News is larger than the ABI in Twitter GloVe pretrained embedding. On the other hand, negative results uncover those attributes whose ABIs are higher in Twitter. In fact, by taking the difference between ABIs in the different embeddings, we are calculating a comparative ABI (CABI), resulting from the use of different informational bases, and we are able to see which embedding includes higher bias for specific attributes. In Figure 4, evidence shows that for attributes related to Allport’s category of “belief” (see Table 2), such as “substandard”, “mediocre” or “inferior” the CABIs are positive, that is, the bias against the poor is relatively higher in Google News Word2Vec than in Twitter GloVe pretrained embeddings. This finding was unexpected in the study, since most sources in Google News are journalists and professionals (Bolukbasi et al. 2016), as compared to Twitter. Although more evidence is needed, this preliminary results could suggest that news could show higher bias against the poor for the attributes that express beliefs.

On the other hand, negative CABIs suggest that bias against the poor is higher in Twitter GloVe, as compared to Google News Word2Vec, when the attributes correspond to Allport’s (1954)

categories of “discrimination” or “physical attack” (see Table 2 in the appendix), that is for attributes such as “hate speech”, “aversion”, “rejection”, “insult” and “contempt”. We find a similar trend, although not as consistent, when comparing the ABIs of unfavourable attributes between Google News Word2Vec and the Wikipedia Glove pretrained embeddings (Figure 5), suggesting that there is higher degree of bias against the poor in Google News in for attributes that express beliefs. When comparing Twitter GloVe and Wikipedia GloVe pretrained embeddings (Figure 6), bias expressed as actions under the categories “discrimination” and even “physical attack” (Table 2) appears to be higher in Twitter, whereas bias expressed as beliefs is higher in Wikipedia or equidistant in the two pretrained embeddings.

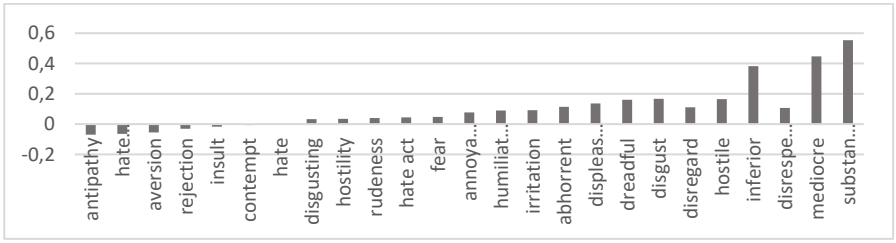

Figure 4: CABIs for unfavourable attributes in Google News Word2Vec vs Twitter GloVe, indicating the difference in the degree of bias per attribute between the two predefined embeddings. Source: authors' creation.

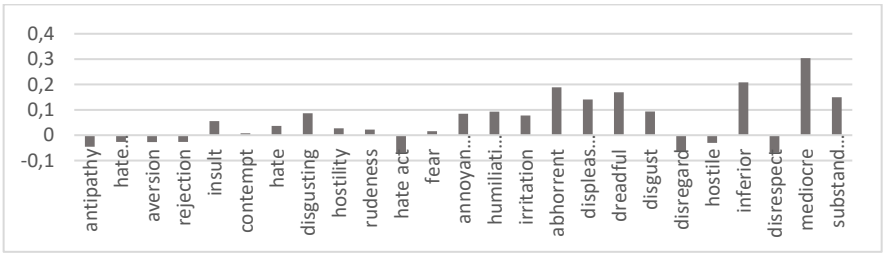

Figure 5: CABIs for unfavourable attributes in Google News vs Wikipedia, indicating the difference in the degree of bias per attribute between the two predefined embeddings. Source: author’s creation

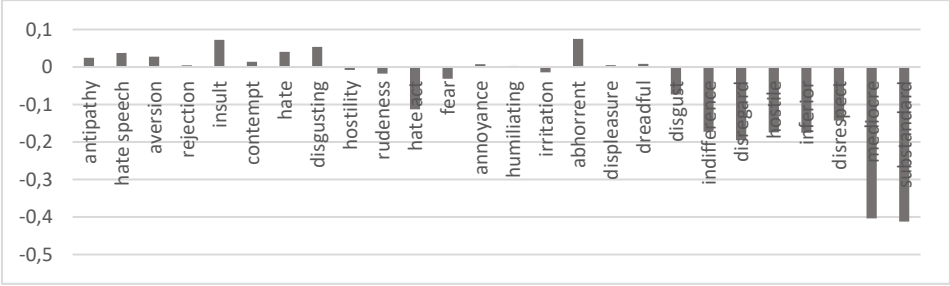

Figure 6: CABIs for unfavourable attributes in Twitter vs Wikipedia, indicating the difference in the degree of bias per attribute between the two predefined embeddings. Source: author's creation

Finally, following Nadeem et al. (2020), we have calculated the distance between the key attributes "rich" and "poor" and neutral attributes using the names of plants, animals and planets, among other terms, in order to test the robustness of the AWEAT model. Although all terms show a bias (that is appear slightly closer to either "rich" or "poor"), only 4 "neutral" terms out of 166 show an ABI level in the order of the first decimal. This proves, on the one hand, that we live in a market economy and therefore all terms have an economic association either to "rich" or "poor". On the other, since this association is much lower than the "favourable" and "unfavourable" attributes used in the study, the test with "neutral" words validates the AWEAT model to evaluate bias against the poor in pretrained embeddings by measuring the distances between "favourable" and "unfavourable" attributes associated to the poor as compared to the rich.

4.5. Conclusions

This study offers a contribution to the body of work on bias with the first set of empirical data evidencing the existence of bias against the poor within the three pretrained word embeddings included in the study, namely Google Word2Vec, Twitter and Wikipedia GloVe. As a result, this paper empirically illustrates a transversal type of bias that has been unnoticed and is aggravated as part of fundamental shared values in welfare states: the belief of equal opportunity and individual responsibility to climb up the ladder. However, when this bias leads inevitably to

discriminatory acts, it has serious consequences towards the achievement of the first Sustainable Goal of the United Nations (no poverty). The article also provides evidence that there is a consistently higher degree of bias in Google News Word2Vec, as compared to the other two embeddings, when the attribute terms express beliefs. On the other hand, a higher level of bias against the poor in Twitter GloVe when the terms express behaviour. These preliminary results could suggest that some news in the media would express a higher level of bias against the poor than individuals in terms of expressed beliefs, whether individuals would offer a higher level of bias shown as behaviour (discrimination or physical attack), for the terms included in the study.

AI systems act as a warning flag of inconspicuous prejudices expressed as bias, but also contribute to spread biased opinions that can eventually lead to discriminatory behaviours. Further studies should be carried out with wider sample of attributes and pretrained embeddings to obtain evidence on the impact of the bias against the poor on the communities that are historically discriminated as a result of other factors, such as gender, race, nationality or religion, to name some examples. A comparative study between the bias against the poor in Global North and the Global South would also be recommended, exploring the correlation between the bias against the poor in line with poverty and inequality levels as well as cultural factors. A deeper analysis is also required to compare biases through different social networks communication channels. Although it is not possible to make the world a better place only through algorithms, they can contribute to make a diagnosis and monitor bias and discriminatory behaviours such as hate speech. This study, therefore, constitutes a first step towards taking action to mitigate pre-existing prejudices that can derive in discriminatory actions. In addition, this work constitutes an evidence for the need to oversee AI technologies and the opportunity that human-in-the-loop decision making, the agreement on pro-ethical development and the implication of social science experts to analyse the roots of bias constitute to convert AI tools not only on autonomous reproducers (and often aggravators) of social inequalities, but on enables for sustainable development.

Appendices of the chapter

Table 1: Proximities and distances between unfavourable attributes and the key terms “poor” and “rich” and the ABI in Google News Word2vec pretrained embeddings. Source: author’s creation

Negative attributes	Proximity to “poor” (cosine)	Proximity to “rich” (cosine)	Relative value: 1 suggests attribute closer to “poor”	Relative distance to “poor” (in radians)	Relative distance to “rich” (in radians)	Aporophobia bias indicator (ABI)
substandard	0,518799	0,065894	1	1,025350	1,504854	0,479503
dreadful	0,496364	0,108623	1	1,051390	1,461958	0,410568
mediocre	0,525181	0,157387	1	1,017868	1,412751	0,394883
inferior	0,442338	0,154269	1	1,11259	1,415908	0,303316
indifference	0,295424	0,049471	1	1,270896	1,521304	0,250408
displeasure	0,181486	-0,043921	1	1,388298	1,614732	0,226433
humiliating	0,236273	0,013788	1	1,332267	1,557007	0,224740
abhorrent	0,177211	-0,034837	1	1,392643	1,605641	0,212997
disgust	0,175618	-0,033866	1	1,394262	1,604669	0,210406
disrespect	0,178972	-0,002676	1	1,390853	1,573472	0,182618
disregard	0,165259	-0,011534	1	1,404775	1,582331	0,177555
fear	0,174980	0,019890	1	1,394910	1,550904	0,155994
irritation	0,152907	0,011789	1	1,417287	1,559006	0,141719
hostile	0,185884	0,045462	1	1,383824	1,525318	0,141493
rudeness	0,176455	0,038615	1	1,393411	1,532171	0,138759
annoyance	0,110991	-0,026991	1	1,459575	1,597791	0,138215
disgusting	0,259967	0,133528	1	1,307807	1,436867	0,129059
hostility	0,132259	0,040978	1	1,438148	1,529806	0,091657
rejection	0,100165	0,037907	1	1,470462	1,532879	0,062416
contempt	0,091754	0,034602	1	1,478912	1,536186	0,057273
hate	0,166657	0,111664	1	1,403357	1,458898	0,055540
insult	0,150543	0,107800	1	1,419678	1,462786	0,043107
aversion	0,169729	0,132875	1	1,400240	1,437526	0,037285
hate act	0,143041	0,111930	1	1,427262	1,458631	0,031369
hate speech	0,154789	0,134926	1	1,415381	1,435456	0,020075
antipathy	0,082810	0,075422	1	1,487891	1,495302	0,007411

Table 2: terms included in the study categorized according to Allport's (1957) degree of action associated to prejudice. Original expressions used by Cortina (2017) appear underlined. Source: author's creation

	Favourable	Unfavourable	
Belief	superior, willpower, kind, courageous, calm, calmness, mildness, mild, innocuous, positive, dignified, delight, delightful, friend, friendship, courage, serenity, excellent, partner, pleasant, polite, brave, higher, adequate, true, happy, peace, peaceful, (28)	Belief	inferior, mediocre, negative, rude, rudeness, lower, shame, shameful, shameless, substandard, slight, carelessness, unkind, inoffensive, distaste, repugnant, rival, scared, sicken, upset, adversary, enemy, opponent (23)
Attitude & action:		Attitude & Action:	
Communication	acknowledgement, empathy, patience, tolerate, attentiveness, respectful speech, patience, cordiality, agreement, endorsement, attestation, regard, taste, remember, interest, tolerance, contentment, politeness, (19)	Antilocution	antipathy, disregard, no acknowledgement, denounce, denunciation, belligerence, belligerent, concern, denial, disagreement, derision, disregard, forget, ignore, indifference, absence of sympathy, refusal, defence, apathy, antagonism (20)
Acceptance	friendliness, friend, goodwill, kind, kindness, sympathy, acceptance, companionable, conciliate, fearless, cordiality, amicability, accord, self-assurance, attraction, desire, recommend, consonance, pleasure, pleasing, confidence, friendly, amity, affability, affection, benevolence, preservation, acquiescence, appetency, liking, becoming, pleasing, solace, love, love speech, liked, acceptance, accept, acceptation, like, complimentary, gentleness, attraction, attractive, approve, approval (46)	Avoidance	disgust, fear, impatience, afraid, alarmed, annoyance, annoying, anxiety, bitterness, challenger, corrupting, defence, defend, detestation, dislike, disgusting, disgust, disapprove, disapproval, detestation, displeasure, dread, dreadful, foe, ill feeling, ill will, irritating, irritation, loathe, loathing, opposition, repel, repugnance, repulse, repulsion, repulsive, resent, resentment, resistance, revulsion, unbecoming, undignified, upsetting, worry, calmness, independence, weighty, hate, abhorrence, abhorrent, hostile, hostility, neglect, unfriendliness, animosity, contempt (56)
Admiration	admiration, praise, approval, appreciation, delight, cherish, adore, flattery, pride, admirable, adulation, praise, dignified, appreciation, appreciate, respect, (16)	Discrimination	degrading, rejection, affront, anger, animosity, aversion, conflict, degrading, demeaning, disrespect, enmity, hatred, intolerance, obstruction, offense, offend, offensive, scorn, slur, shamed, unsupportive, hostility, abandonment, humiliating, hate speech, insult (27)
Aid	Aid, help, heal, support, love act, cooperation, comfort, facilitation, ally, shelter, encourage, encouraging (12),	Physical attack	hate act, physical aggression, abuse, abusive, aggression, assault, attack, bellicose, bellicosity, intimidate, intimidating, intimidation, violence, violent, harm, physical protection (16)

Table 3: Proximities and distances between favourable attributes and the key terms “poor” and “rich” and the ABI in Google News Word2vec pretrained embeddings

Favourable attributes	Proximity to “poor” (cosine)	Proximity to “rich” (cosine)	Relative value: 1 suggests attribute closer to the poor	Relative distance to “poor” (in radians)	Relative distance to “rich” (in radians)	Aporophobia bias indicator (ABI)
sympathy	0,169531	0,018321	1	1,400441	1,552474	0,152032
politeness	0,132293	0,068439	1	1,438114	1,502303	0,064189
pleasing	0,227241	0,174897	1	1,341551	1,394995	0,053443
goodwill	0,088890	0,039868	1	1,481787	1,530918	0,049129
cordiality	0,043623	0,007792	1	1,527159	1,563004	0,035845
happy	0,212202	0,180576	1	1,356968	1,389223	0,032255
fearless	0,100959	0,069186	1	1,469664	1,501554	0,031889
pride	0,104457	0,088019	1	1,466148	1,482663	0,016514
friendliness	0,178084	0,175157	1	1,391756	1,394731	0,002974
courageous	1	1	0	0	0	0
self-assurance	1	1	0	0	0	0
carelessness	1	1	0	0	0	0
defence	1	1	0	0	0	0
affection	0,100301	0,10674	0	1,470325	1,463852	-0,006474
liked	0,125296	0,135883	0	1,445169	1,434491	-0,010678
delight	0,033640	0,045317	0	1,537149	1,525463	-0,011687
desire	0,085015	0,096916	0	1,485677	1,473728	-0,011949
pleasant	0,168783	0,187770	0	1,401201	1,381905	-0,019297
acceptation	0,049464	0,099845	0	1,521311	1,470784	-0,050527
appreciation	0,005268	0,075830	0	1,565527	1,494893	-0,070635
independence	0,067198	0,141933	0	1,503546	1,428382	-0,075165
love	0,107482	0,184401	0	1,463105	1,385334	-0,077772
delightful	0,131124	0,215119	0	1,439293	1,353983	-0,085311
flattery	0,054658	0,140086	0	1,516110	1,430247	-0,085864
friendly	0,184168	0,271432	0	1,385570	1,295916	-0,089655
endorsement	-0,049720	0,057279	0	1,620537	1,513486	-0,107052
taste	0,147377	0,261997	0	1,422879	1,305705	-0,117175
pleasure	-0,005007	0,120311	0	1,575803	1,450193	-0,125610

Chapter 5. From ethical principles to practices: a hands-on process to manage bias in the design of NLP systems

Summary of the chapter

The Artificial Intelligence (AI) legal framework being implemented in the European Union incorporates the principle of diversity, non-discrimination and fairness, including the avoidance of unfair bias, which has been dealt from the data and algorithmic point of view. However, the phenomenon of bias can also be tackled as a pro-ethical process in the development of AI Natural Language Processing (NLP) systems. This study provides a conceptual background clarifying the concepts of prejudices, discrimination, stereotypes and bias and explains the aggravating particularities of bias in NLP systems. It identifies the gaps in the existing literature to assist AI teams in the management of bias within the development process. Then, based on a strong ethical framework, the paper proposes an end-to-end procedure that translates the non-discrimination ethical principle into specific actions to be applied within each step of the design, building, testing, deployment and monitoring of an NLP system. The paper aims to support AI development teams to identify, mitigate and monitor bias in practice, with the involvement of target users, including the historically discriminated groups. Additionally, the paper provides guidance on how to explain the always imperfect trade-offs in terms of bias to users, since a system that is explainable is inherently fairer.

Keywords: Bias, Artificial Intelligence, Trustworthy AI, fairness, discrimination, pro-ethical design.

5.1. Introduction

Discrimination and bias in AI are key topics in the current information civilization (Zuboff 2019), where behavioural data feeds AI systems that influence human behaviour. The European Commission expects that by 2025 the economic impact of AI will reach between 6.5 and 12

trillion euro annually (European Comission 2019) and governmental efforts to regulate AI have gained traction in the past few years. In particular, in the grounds of domestic AI governance, the European Union is considered to have an ethically superior legal framework being put in place as regards citizens' rights (Allison and Schmidt 2020; Gill 2020; Imbrie et al. 2020; Roberts et al. 2021) and its repercussion reaches well across the EU borders since companies in other countries will tend to comply with EU regulations in order to have a single goal approach (Bradford 2020). A fundamental piece of the EU AI framework, described in the Ethics Guidelines for Trustworthy AI (HLEGAI 2019) and the proposed EU Artificial Intelligence Act (European Commission 2021), is the concept human-centric trustworthy AI, which contains the principle of "diversity, non-discrimination and fairness", including the "avoidance of unfair bias", especially in the case of the historically discriminated groups (HDG) (HLEGAI 2019). However, the human-centred approach defined by the AI Act has immediate criticisms. On the one hand, it has been criticised as anthropocentric, suggesting that it neglects environmental problems (Floridi 2021) and it does not incorporate the new trend of "green" AI (Schwartz et al. 2019; Cowls et al. 2021). On the other hand, it has been stated that the Trustworthy AI principles have not been defined for intrinsically ethical reasons, since the AI Act has a risk-based approach which is commonly used to provide the required public trust in the market (Floridi et al. 2018; Fuster and Brussel 2020; Roberts et al. 2021). In addition, the AI Act has been qualified as ambiguous, because it proposes a mix of hard and soft approaches to ethical principles and there is confusion about when it is compulsory or not to comply with the rule. Other authors consider that the AI Act does not actively incentivise AI for social good in terms of contributing to meet the United Nations Sustainable Development Goals (Vinuesa et al. 2020; Roberts et al. 2021).

But most of all, it has been argued that the EU Trustworthy AI principles, although appropriate, are abstract and AI development teams find them difficult to apply in practice, highlighting the urgent need to go from the "what" to the "how" and translate these principles into practice (Floridi 2019b; Vakkuri et al. 2020; Morley et al. 2021a, b). This pitfall, however, is not exclusive of the

EU AI legal framework, but to the nascent field of AI ethics in general (Vakkuri and Kemell 2019). Back in the mid 2019, Jobin et al identified more than 80 ethics guides available for public domain (Jobin et al. 2019) and the Global Inventory of AI Ethics Guidelines identified more than 173 documents in existence in 2021 (Algorithm 2021). These documents also focus the effort on the "what" and fall short to clarify the operationalization and empirical evaluation of AI ethics (Peters 2019; Morley et al. 2021a). Given this context, this paper aims to answer the question: how can the principle of "non-discrimination and fairness" be applied into the practice of the AI development? Although this topic has been tackled from a data and algorithmic point of view, it has been suggested that these are often *ad hoc* narrow approaches and that a more holistic approach is required, with focus on procedural regularity (Morley et al. 2021a), well-grounded on a conceptual framework that explains the causes of bias not only in AI, but also in the off-line, since bias and discrimination are not exclusive of the on-line world (Blodgett et al. 2020; Card and Smith 2020).

It is analysing the general nature of bias, based ultimately on our beliefs, where it becomes clear that we cannot draw a hard line between what is sufficient and insufficient proof of discrimination and how to avoid it (Allport 1954). Therefore, the way to implement the "non-discrimination principle" seems to be by pursuing fairness as "an appropriate concession" and manage rather than seek to completely mitigate bias, understanding the trade-offs in specific backgrounds and state of affairs (van Nood and Yeomans 2021). In this context, the three axes defined by Floridi & Taddeo: "ethics of data", "ethics of algorithms" and "ethics of practices" (2016), highlight the need to manage AI bias also from a process design point of view, incorporating the need to mitigate, whenever possible, bias in data and algorithms, since the three axes are intertwined. In order to do that, first of all, this article provides a robust conceptual framework to describe the nature of prejudice, discrimination, stereotypes and bias both in the online and the offline. Then, we explain the aggravating factors of bias in AI NLP tools and review the state-of-the-art approaches to bias from an algorithmic and data perspectives. From there, we propose a hands-

on end-to-end process to assist AI development teams to manage bias at each step of the design, building, testing, deployment and monitoring of an NLP system, where responsibility is distributed across different agents since there is an active participation of target users including historically discriminated groups. As a result, the process described facilitates AI practitioners to disclose the always imperfect trade-offs. Finally, we conclude explaining actions that need to be taken to evaluate the process and expand pro-ethics design with focus on procedures to other AI Trustworthy principles and AI branches.

5.2. Conceptual framework: what is discrimination, prejudice and bias in general terms

After analysing 146 papers studying bias in NLP systems (published prior to May 2020), Blodgett et al (2020) concluded that quantitative techniques to measure or mitigate biases are poorly matched to their motivations and often there are self-evident statements of bias, since these papers do not provide an actual conceptualization of bias outside NLP systems. Card & Smith (2020) state that literature on fairness within machine learning largely depends on assumptions. In order to analyse how to deal with bias in AI it is important to understand first what are the causes of the discriminatory phenomena in the "onlife", using Floridi's term (2015). This paper fills in the existing conceptualization gap in the AI bias literature.

Discrimination has been widely documented in literature as an action of exclusion resulting from prejudice, which Allport (1954: 7) defines as "an avertive or hostile attitude toward a person who belongs to a group, simply because he belongs to that group, and therefore presumed to have the objectionable qualities ascribed to the group".

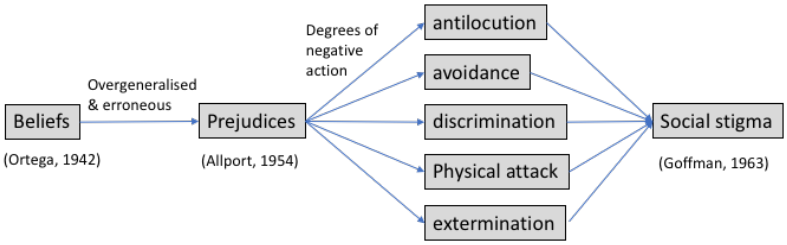

Figure 1: prejudices are overgeneralised and erroneous beliefs that can be classified according to the degree of action (Allport 1954) and generate social stigma (Goffman 1963). Source: author's creation

Prejudices can be favourable or unfavourable and can bring different degrees of action defined by Allport as antilocution, avoidance, discrimination, physical attack and even extermination (Fig 1). In this context, it has been suggested that the resulting action of prejudice can be considered discrimination when "the prejudiced person makes detrimental distinctions of an active sort" (1954: 14). Unfavourable discrimination brings a social stigma, which is associated with feelings of shame on the side of the discriminated (Goffman 1963) and beliefs of deservingness on the side of the discriminators, for instance in the case of prejudice resulting from socio-economic factors (Arneson 1997; Applebaum 2001; Everatt 2009; Nunn and Biressi 2009). When prejudices have a social category, we are talking about stereotypes, which are transmitted through the linguistic process, what we know as bias, creating a self-perpetuating cycle where prejudices are shared and maintained (Maass 1999; Beukeboom and Burgers 2019) (Fig 2). Therefore, bias is the linguistic expression of shared social prejudices within a specific culture.

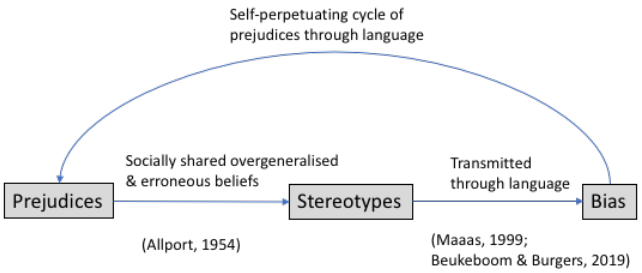

Figure 2: when prejudices are shared within a specific culture we can talk about stereotypes and bias is the transmission of stereotypes through language, which creates a vicious cycle self-perpetuating prejudices. Source: author's creation

It becomes clear that in order to know how to deal with the Trustworthy AI principle of diversity, non-discrimination and fairness, including avoidance of unfair bias, we need to analyse the nature of prejudices which are at the root of bias and other discriminatory phenomena. In this respect, Allport (1954) explains that the reason why human beings are not solely evaluated based on their individual characteristics is because there is not enough time to understand every single object, person or phenomenon in detail in order to make the necessary decisions to survive. Therefore, human beings put information into categories and generalise based on the previous experience. For example, a child learns that a tree has some roots, a trunk, branches and leaves. That is enough to identify a tree and there is no need to know all the different species included in books on botanic science. According to Allport, few human judgements are based on absolute certainty, but on some probability based on the categorization and previous experience of the individual. Therefore, we cannot draw a hard line between what is and what is not a prejudice or bias and we can say that all humans are discriminated and discriminators. Although prejudices and bias are difficult to identify, can they be mitigated? Allport suggests that prejudices are overgeneralized (and therefore erroneous beliefs) that lead to an attitude of favour or disfavour (Allport 1954). Allport, however, does not define beliefs. It is in the work of Ortega y Gasset (1942) where we find that beliefs are the framework that allows us to interpret the world we live in. Beliefs are part of ourselves and the only way to question them is to become aware of them, transforming them into ideas, which is only possible when acquiring new knowledge that allows for critical thought and empathy, overcoming the emotional resistance linked to prejudices (Morgado 2017). However, developing the critical thought that counteracts prejudices is often a complex endeavour, since discrimination tends to be a multifactorial phenomenon, where different types of prejudice tangle and aggravate one another. For example, the prejudices that faces a refugee poor woman that belongs to an ethnic minority are not the same as a rich woman in that same situation, or even less if the person belongs to the same ethnic group or to the same country.

Another way to tackle bias and prejudices is to analyse the concept of fairness, which is the other side of the coin and has been traditionally defined as "giving to each person what corresponds to him or her" (Cortina 2007: chapter 6.7) or according to Rawls (1971), as "distributive justice", understood as the best possible distribution for the least advantaged members of society as long as this is consistent with the freedom of all. As it happens with complete bias mitigation, complete fairness has never been achieved (Cortina 2007), since it is subjective and therefore dependable on each culture and even person. But how is this translated in the practice of AI NLP development? Some authors defend that AI models should be tailored for the values of the societies where they operate (Carman and Rosman 2021), others consider working towards an intercultural citizenship and universal values (Jiang et al. 2021), which seems a feasible objective considering the cohesion and communication alternatives that new technologies offer. There is also the approach of the ethical pluralism (Ess 2020; Wong 2020) which acknowledges the coexistence of universally valid values that might be interpreted differently across a diversity of cultures and constitutes a challenge for AI systems. More specific methods to achieve fairness in algorithmic systems can be found under the mathematical approach. These methods can be listed as: fairness through unawareness (Hardt et al. 2016; Card and Smith 2020), demographic or statistical parity (Dwork et al. 2011), individual fairness (Green and Hu 2018), randomisation (Kroll et al. 2017b), equality of Odds / Opportunity (Hardt et al. 2016). However, all these mathematical approximations to the concept of fairness have drawbacks and are mutually incompatible (Kleinberg et al. 2016; Card and Smith 2020; Tsamados et al. 2021b). Acknowledging the incapability to achieve complete fairness and to deal with it as an appropriate concession, which requires trade-offs, can help AI practitioners explain decisions openly to users (van Nood and Yeomans 2021). Either if we look at the Trustworthy AI principle from the prejudices or from the fairness perspective, we realise that there are no absolute black or white solutions, rather the way forward seems to be to manage bias by agreeing trade-offs with stakeholders, bearing in mind that we are dealing with abstract concepts that evolve over time

(MacMahon 2016) in line with global issues (van Nood and Yeomans 2021). In this respect, it is essential and urgent to provide AI development teams with tools that incorporate the three axes: data, algorithm and practices (Floridi and Taddeo 2016), where the latter in particular becomes highly relevant when we acknowledge the continuously improving process towards fairness and mitigation of unfair bias is a goal in itself.

5.3. Bias in NLP Systems

AI systems need data not only to work but also to train how to do the work. The need for Big Data, makes quality, ethical standards and relevance of data are very hard to assure, inheriting bias in society, which is part of natural language (Rudinger et al. 2018; Chiappa et al. 2020). Although the emphasis is changing from the quantity to the quality of "greener" data sets (Schick and Schütze 2020) improving performance of data models (Schick and Schütze 2020), and even using synthetic data as opposed to historical data (Watson et al. 2019) which can be aligned according to specific value systems to manage bias (Sierra et al. 2021) always partial in the sense of being incomplete (van Nood and Yeomans 2021). Algorithms can also be biased on the way they learn or, more appropriately, on the way they are programmed to learn (Mittelstadt et al.; Tsamados et al. 2021b). Additionally, AI solutions are deployed into real complex systems, where agents interact, making it difficult to predict the social impact of an algorithmic system before actually deploying it (Morley et al. 2020). Finally, biases can have their origin in practices. For example, due to the high volume of data involved and the data volatility, analysts themselves introduce bias on data gathering by reducing the scope and therefore compromising the veracity of the data and therefore the AI models to extract it (De Mauro et al. 2016; Martínez-Plumed et al. 2019). But, most importantly, unknown bias in algorithmic systems can also be the result of not having defined procedures to manage and deal with bias, which would allow for conscious and agreed decisions with stakeholders (Floridi 2019a; Vakkuri et al. 2020) Although there are initial reviews that aim to translate Trustworthy AI principles into practices (Morley et al. 2020),

to date there is not an end-to-end process describing how to mitigate AI bias throughout the different development stages. This article aims to fill in this gap.

5.4. The effects of bias in NLP systems

Since part of AI bias is originated from socially-shared prejudices, should we not consider that finding bias in Artificial Intelligence systems is normal? After all, society itself communicates in a biased way…But bias in AI, in NLP models in particular, has some particularities that deserve special attention. To start with, users are not aware of it, which is a key topic since, as we have described, knowledge contributes to mitigate prejudice and discrimination. The lack of transparency is aggravated by the fact that there is a generalised over-trusting or even blind faith in AI, illustrated in the case of the man who fell from a cliff in his car following the GPS car navigation system and admitted that "he didn't think to over-rule the machine's instructions" (Fry 2018 : 16). To make things worse, there is also a lack of accountability for discrimination, as described by Barocas and Selbst: "discrimination is almost always an unintentional emergent property of the algorithm's use rather than a conscious choice by its programmers, it can be unusually hard to identify the source of the problem or to explain it to a court" (Barocas and Selbst 2016 : 1). Most importantly, bias is often amplified in the use of AI models. This is the case because word embeddings represent words by vectors (suitable to be processed by machine learning algorithms). For example, as a result of bias in embeddings, we find that names associated to one gender appear closer to specific professions (or example, "John" is closer to computer programmer than "Mary". On the other hand, "Mary" is closer to "homemaker"). Such bias in word embeddings amplifies the already existing bias in society because when searching for computers programmers on the Internet, "John"'s page will appear first, making even harder for women to be recognised as computer programmers (Bolukbasi et al. 2016). This is a serious topic since pretrained word embeddings, which are publicly available and easy to incorporate in apps, are used in a wide range of domains, such as financial lending, personnel hiring, targeting of marketing campaigns and even critical sectors such as health and justice, with the resulting

social impact (Solaiman et al. 2019). However, the detection of biases in AI systems also have positive consequences: biases in AI models trained on historical data reflects bias in society, therefore act as a mirror, showing the prejudices that go unnoticed off-line. Since knowledge on prejudices (Allport 1954) and empathy (Cortina 2007) are ways forward to mitigate bias, detecting, measuring and following up bias in AI systems can contribute to the improvement of equity and fairness in both the digital and the tangible world.

5.5. Detecting and de-biasing data

A growing number of studies analyses how to detect bias in data using a diversity of methodologies. Kiritchenko et al (2014) use 219 automatic sentiment analysis systems to find out if there is consistent higher sentiment predictions for one race or gender, by rating the emotion expressed by two sentences that only differ on the gender or race of the person mentioned (such as “this man made me feel angry” vs “this woman made me feel angry”). Word analogy is another popular method for testing bias in word embeddings: given two pair of words in a certain syntactic or semantic relation (man : king) and (woman : queen), the goal is to find out if embeddings capture gender bias by semantic relations (such as doctor : man :: woman : nurse) (Bolukbasi et al. 2016). Following this line, Nadeem et al (2020) measures stereotypical bias in popular pretrained language models (BERT, Roberta, XLNet and GPT2) with respect to their language ability. In their study, bias is measured for the domains of gender, profession, race and religion within the US geographical area using both intrasentence and intersentence associations, concluding that as the language model becomes stronger, stereotypical bias too. In other words, the more human-like are the word embeddings, the more bias can be found. Other studies (Nikhil Garga, Londa Schiebingerb, Dan Jurafskyc 2018) provide the tracking of bias over time, which allows us to relate bias with global events such as terrorist attacks or the percentage of women occupation in certain professions. Bolukbasi et al (2016) presented results for debiasing gender stereotypes for the 50.000 most frequently used words in Word2Vec (trained on Google News – 3M English words). It is interesting to highlight that not only direct bias is considered, identified

with word analogies such as “woman: nurse :: man : doctor”, but also indirect bias such as “receptionist is closer to softball than to football”. Other studies perform a multiclass debiasing including social class and religion (Manzini et al. 2019). More recently, Zhao et al (2021) presented the Linguistic Ethical Interventions (LEI) to mitigate bias by communicating context specific principles of ethics and equity.

5.6. Training algorithms to detect bias

Most recently, approaches have been presented to train algorithms with moral judgements material so that they can detect bias. Sap et al (2020) developed a model that incorporates annotations aiming to consider common-sense reasoning on social implications. The model has been trained in the “Social Bias Inference Corpus” which contains 150 K structured annotations of social media posts covering 34K implications about demographic groups, including sources such as English Twitter datasets annotated for toxic or abusive language. In turn, researchers of the Dephi project (Jiang et al. 2021), have built a prototype that aims to explicitly train state-of-the-art AI models on moral judgements, weighing competing moral concerns and conflicts between broad ethical norms and personal values. The “Commonsense Norm Bank” of the Dephi project compiles 1,7 M examples of people judgements of various real-world situations. However, since it is built out of existing judgements datasets, the “original” Delphi made mistakes in terms of social biases and inequality, perpetuating racism and sexism 9% and 3% of the time. As a result, the model had to be corrected to reduce bias to 2%.

5.7. The missing axe in the management of bias

Despite the existing data and algorithmic approaches that deal with bias and the proliferation of principle-based AI ethic guides (Morley et al. 2021a), 79% of tech workers admit that they would like practical resources to assist them with ethical considerations (Miller and Coldicott 2019). Therefore, there is a gap between theory and practice in the AI ethics field (Vakkuri and Kemell 2019). On the other hand, data and algorithmic approaches, although necessary, constitute narrow

and insufficient approaches to manage an evolving complex ethical topic where trade-offs need to be agreed among stakeholders. The axe defined as "practices" by Floridi and Taddeo (2016) is the missing and required piece. To add pressure to AI practitioners, the Trustworthy AI principles will be compulsory by law once the EU AI act is enacted which, in the current lack of clarity, contributes to creating "moral panic" (Ess 2020) where counterproductive practices such as ethics shopping, ethics bluewashing, ethics lobbying, ethics dumping or ethics shirking can flourish (Floridi 2019a). In other words, it has been claimed that there is an urgency to translate theoretical principles into practice (Morley et al. 2021a) by defining inclusive processes that can be carried out by multidisciplinary teams, facilitating public involvement in decisions concerning fairness (Harrison et al. 2020). Although pro-ethical design can imply some overheads, especially if it means modifying standard practices, it is also recognised that it improves social impact, consumer trust and satisfaction, public reputation and it reassures investors (Morley et al. 2021b). Therefore, pro-ethical design of AI systems will become easier to introduce as societies where it operates are better informed, become more critically aware and mature on the topic of digital ethics so that there is an actual demand for it. The legal framework being implemented on trustworthy AI will contribute to that end.

5.8. A process to bring the non-discrimination principle down to the design level

This chapter aims at assisting NLP development teams in the implementation of Trustworthy AI principle of "diversity, non-discrimination and fairness, including avoidance of unfair bias" by providing a hands-on end-to-end process that brings the principle to the design level. The main original contribution of this article, therefore, is the AI Bias Mitigation Process (AIBMP) (Fig 3), which does not aim to constrain the choices of agents, but encourages agents to make choices, in line with the pro-ethical design concept. Compared to the ethics by design approach, pro-ethical design can be considered an "ethical attitude" described by Floridi (2019c), changing the focus from principles to practical decision-making-processes in line with the Aristotelian concept of

phronesis. Morley et al describe a pro-ethical design approach as shifting the focus from a "paternalistic imposition of inflexible standards that ignore context and more about procedural regularity and public reason that can be adapted and shared across contexts and societies" (Morley et al. 2021a : 247). In other words, the AIBMP is not a top-down procedure to set specific norms, but seeks to mitigate bias by involving main stakeholders into a rational argumentation, considering cultural and context specific Machine Learning (ML) ethics. It aims to be a reflective development process that aids AI practitioners understand also their own subjectivity and biases in a specific context, as recommended by Terzis (2020). It is also important to mention that AIBMP aims to be led by a team of multi-disciplinary researchers and proposes the active involvement of a representation of target users (including historically discriminated groups (HDG)), to be able to communicate and reach agreements on the inevitable trade-offs (Morley et al. 2021a). The process, therefore, foresees that target users participate not only expressing their needs and providing feedback, but being invited to participate in co-creation sessions, in line with the user agency concept described in the EU Ethics Guidelines for Trustworthy AI (HLEGAI 2019).

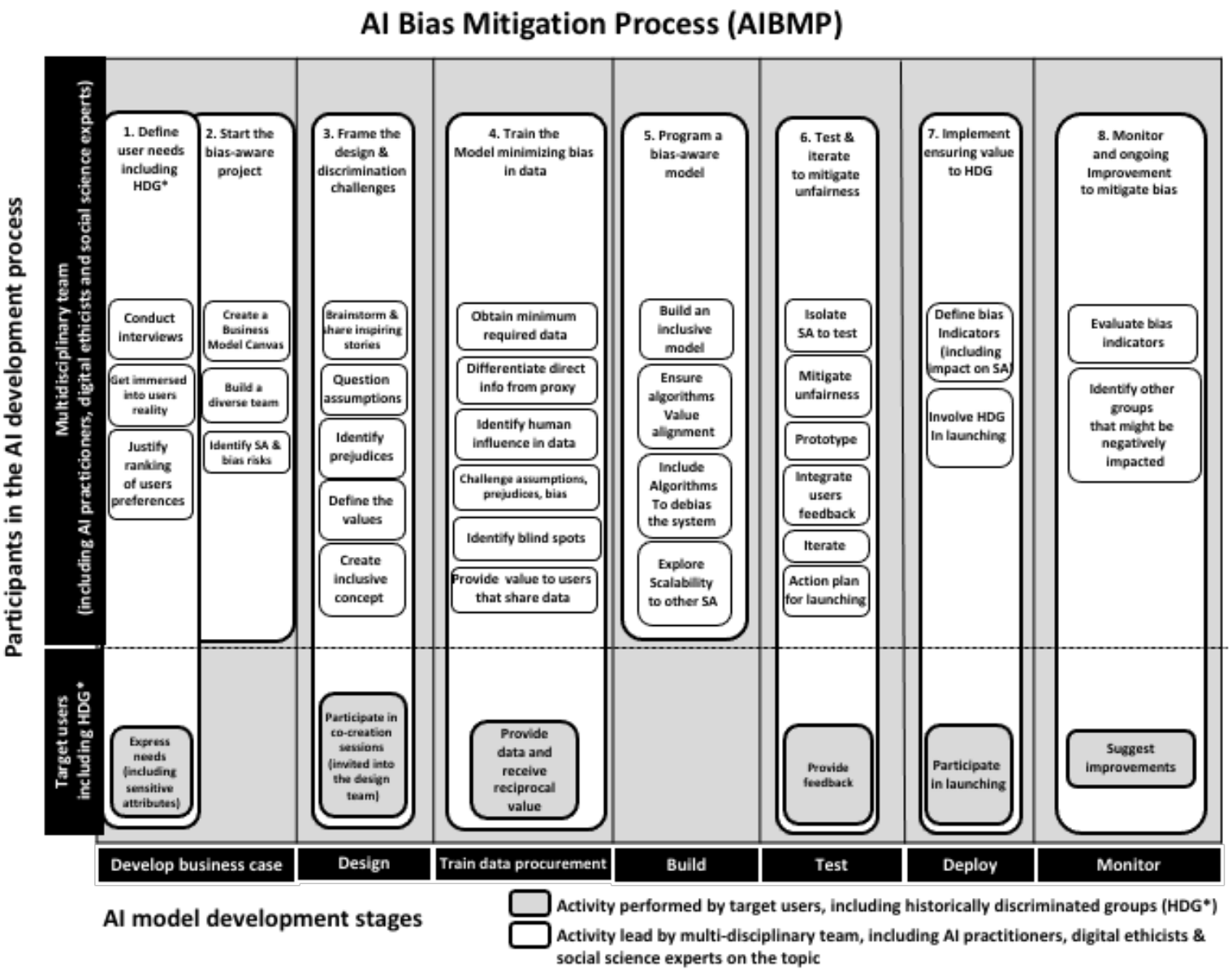

Figure 3: the AI Bias Mitigation Process (AIBMP) translates the principle of diversity, non-discrimination and fairness, including the avoidance of unfair bias (HLEGAI 2019) into the practice of AI NLP development. It is aimed at assisting AI NLP multidisciplinary development teams to understand their own subjectivity and biases and share the responsibility of trade-offs decision making among the different stakeholders. Target users have an active role in the process, including historically discriminated groups (HDG). Source: Author's creation

Clarifications for each step of the AIBMP:

1. Define user needs, including historically discriminated groups (HDG). User needs are the centre of human-centred design (Ideo.org 2015). The deeper the AI Development Multidisciplinary Team (AIDMT) gets into the users' reality (including HDG), the more will be able to understand users' beliefs and values and therefore question social assumptions and prejudices, mitigating bias. As there are always trade-offs in all development processes, AIDMT needs to justify the ranking of users' preferences in order to provide explainability.
2. Start the bias-aware project. Building a diverse team is an integral part of the project in order to achieve ethical pluralism (Ess 2020), bearing in mind international ethical principles and implementing them according to specific culture and context requirements. Indeed, the practical operationalisation of AI ethics is not about external impositions, but more about practical wisdom, in line with Aristotle's concept of *phronesis* or about procedural regularity, continuously learning from own subjectivity and biases, adapting the process across contexts, reaching agreements with stakeholders and within the multidisciplinary team (Kroll et al. 2017b). This team should identify the bias risks and sensitive attributes being taken into account in the business model canvas.
3. Frame the design & discrimination challenges. The needs of the HDG should be taken into account in the brainstorming sessions, paying special attention to inspiring stories, since personalising users provide knowledge on other cultures and contexts that help identifying values, assumptions and counteract prejudices, which are at the origin of bias (Allport 1954). The objective is to create an inclusive concept with the participation of a representation of target users that should be invited into the design team in co-creation sessions (Ideo.org 2015), in line with the "human agency" principle. The design process

should avoid one-size-fits-all approach and consider universal design principles for the widest possible range of users (HLEGAI 2019).

4. Train the model minimizing bias in data. Enormous amounts of data tend to include low quality information, therefore, the minimum amount of quality data should be obtained (Schick and Schütze 2020), which makes bias easier to manage. Once the minimum required quality data is selected, it needs to be analysed to challenge assumptions, prejudices and the resulting bias by differentiating direct information from proxy, identifying human influence in data as well as blind spots (Sampson, O., & Chapman 2021). The existing technical approaches to identify and measure bias in data can be explored (Kiritchenko et al. 2014; Bolukbasi et al. 2016; Nikhil Garga, Londa Schiebingerb, Dan Jurafskyc 2018; Manzini et al. 2019; Nadeem et al. 2020; Zhao et al. 2021). In addition, the AIDMT needs to foresee that value should be provided to users that share data in order to comply with fairness criteria.
5. Program a bias-aware model. Since data from the "real world" cannot be assumed to have the same values, algorithms to debias the system once it is using "uncontrolled" data are to be foreseen (Bolukbasi et al. 2016; Manzini et al. 2019; Zhao et al. 2021). The development team can consider using methods to train algorithms to detect bias (Sap et al. 2020; Jiang et al. 2021). Scalability to other sensitive attributes (SA) needs to be explored in order to enhance the inclusiveness of the model.
6. Test & iterate to mitigate unfairness. Sensitive attributes (SA) are to be tested in isolation to ensure unfairness is mitigated and users' feedback can be integrated into several iterations of the prototype, in line with the concept of non-bias engineering of negotiated ethics (Morley et al. 2021a).
7. Implement ensuring value to HDG. Indicators are to be defined in order to measure and monitor impact of the AI model on SA. This information needs to be publicly available in the launching of the AI model and thereafter in line with the transparency principle. Target users, including HDG, are invited to participate in the launch as recommended by

"stakeholder participation" in the EU Ethics Guidelines for Trustworthy AI (HLEGAI 2019).

8. Monitor and ongoing improvement to mitigate bias. The AIBMP has been defined as an integral part of the AI systems design, development and implementation process, as recommended by Arnold and Scheutz (2018). It is not to be applied as a "one-off" test, but to be re-applied on ongoing basis as AI systems are revised and re-tuned, understanding AI ethics focuses on procedural regularity (Morley et al. 2021a). The agreed values should be revisited and enriched in line with the maturity of society. In other words, the AIBMP aims at avoiding the phenomenon described as ethics by "tick-box" (Morley et al. 2020), since it constitutes a multidisciplinary and ongoing process of reflection, helping AI practitioners to understand their own subjectivity and biases within given circumstances (Terzis 2020), highlighting why unethical results occur so that the appropriate avoidance strategy can be implemented (Fazelpour and Lipton 2020).

The AIBMP, however, does not only intend to assist the AI development teams to manage bias inclusively, but also to create the grounds to be able to disclose the imperfect trade-offs, involved in decisions dealing with bias, to persons, professionals and authorities, allowing them to judge on the limitations and fairness of the system and the possibility to use it or not accordingly. In other words, the AIBMP facilitates the principle of explainability (HLEGAI 2019). Since decisions are reflected upon and agreed among stakeholders, they are easier to communicate. The explainability principle has been described as a second-order principle since it can be directly addressed from a programming perspective avoiding the black-box effect (Floridi et al. 2018; Carman and Rosman 2021), which allows organizations to defend the always imperfect trade-offs (Whittlestone et al. 2019). In fact, it is argued that when a system is explainable and interpretable it is inherently more accountable and fairer (Binns; Fazelpour and Lipton 2020) Since AI systems need to be designed to be transparent from the beginning (Ananny and Crawford 2018), Figure 4

defines what is the minimum information from the AIMBP that needs to be explicitly communicated in order to understand what the AI systems aims to achieve, how they do it and why they do it in that particular way (Kroll 2018).

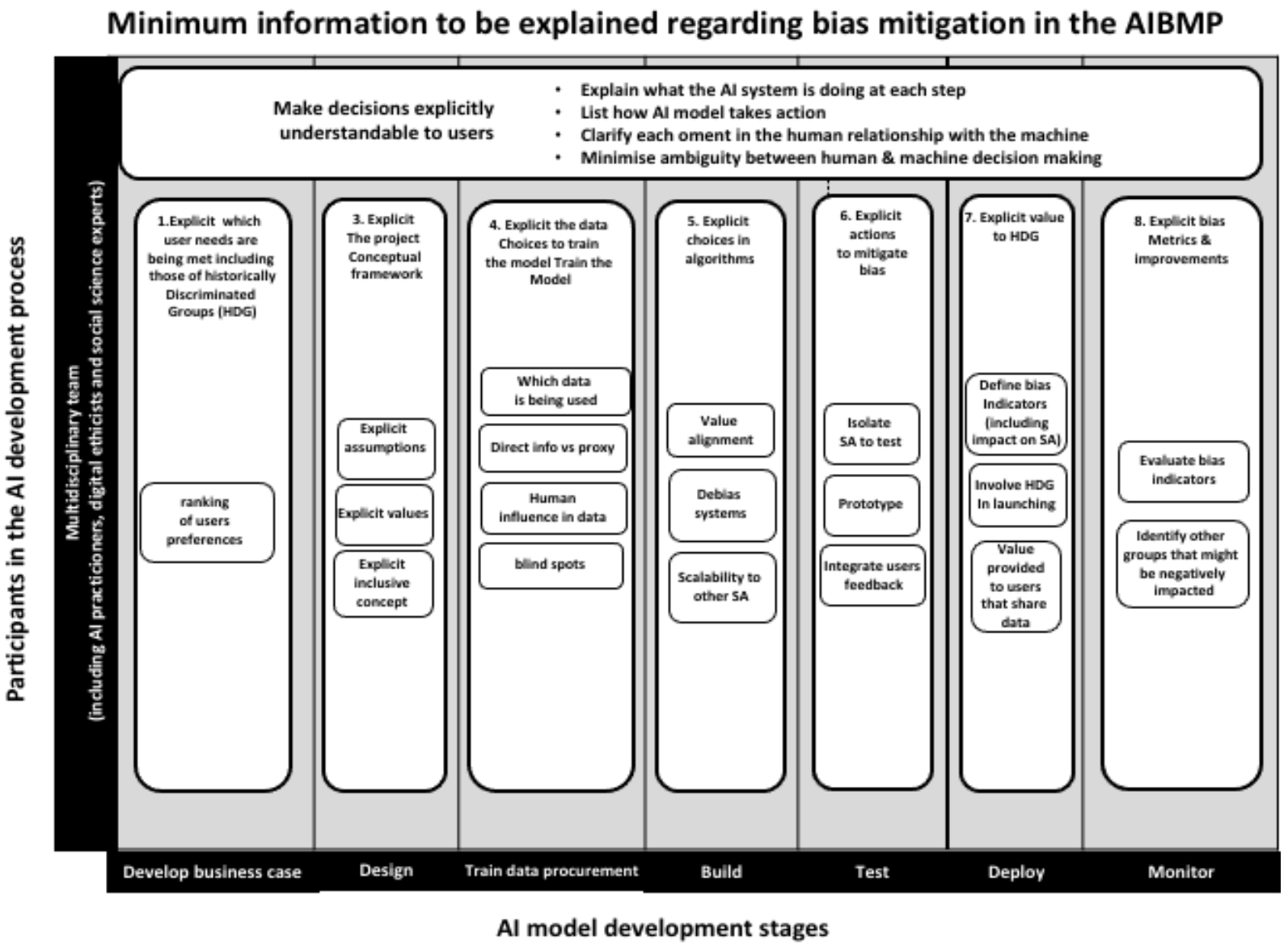

Figure 4: description of the minimum information throughout the AI Bias Mitigation Process that needs to be disclosed so that users can take ethically informed decisions, in line with the explainability principle (HLEGAI 2019). Source: author's creation

All approaches that deal with complex ethical principles have some limitations and this is not an exception. The AIMBP itself has not yet been empirically validated. Therefore, further studies are required to test AIMBP model in a real context, by a multidisciplinary team and real data. Additionally, a multi-disciplinary ethics advisory board, such as the trial performed by Digital Catapult (Morley et al. 2021a), should evaluate the suitability and comprehensiveness of the process. Procedures for compliance and auditing also need to be developed.

5.9. Conclusions

In recent years numerous studies have acknowledged that AI systems, in particular NLP models, can have harmful consequences on the grounds of discrimination, aggravated by a generalised over-trusting of technology. However, there is a lack of transparency and accountability on bias in AI systems, which has serious consequences on human rights since word embeddings are used in critical sectors such as health, education and justice. In this context, there has been a proliferation of principle-based ethic codes. Domestic legal frameworks are also being put in place worldwide to regulate the development of AI systems, with special focus in the European Union on protecting human rights including the principle of diversity, non-discrimination, fairness and the avoidance of unfair bias. However, it remains unclear how to comply with this principle in practical terms when designing an AI model. Although there is a significant body of work describing how to detect and correct biases in NLP systems in the data and algorithms, these are narrow and often *ad hoc* solutions, which are often based on assumptions and fail to actually clarify the nature of "bias" in the first place.

To fill in the existing gaps, this chapter provides, first of all, a descriptive framework for the concepts of prejudice, discrimination, stereotypes and bias. Since prejudices are originated in the way human beings interpret reality, bias cannot be mitigated completely, rather it should be managed not only in the data and algorithms but also in the practices of AI NLP development.

While existing AI literature on bias focus on technical solutions in control environments, the AI Bias Mitigation Process (AIBMP) seeks to provide an end-to-end inclusive framework that encourages an ongoing reflective approach to bias mitigation and management, by suggesting specific actions to be taken by a multi-disciplinary team and active involvement of target users, including HDG. As a result, the AIBMP facilitates the disclosure of the trade-offs when managing bias, providing users with the necessary information to take ethically-informed decisions.

Quality research and empirical testing will be required to prove the applicability of the AIBMP, ensure the benefits on bias mitigation and improve the drawbacks. Procedures for compliance and auditing also need to be developed. Additionally, further work should be performed translating the rest of Trustworthy AI principles into pro-ethics design processes, including specifications for economic sectors and AI branches. Although such processes might be seen as overheads initially, societies where AI systems operate are becoming better informed, more critically aware and mature on the topic of digital ethics. As a result, demand will also encourage the focus on continuous improvement of ethical standards, which cannot be achieved as one-shot activity nor with narrow ad hoc solutions, but rather be the result of procedural regularity and inclusive participation.

Chapter 6. A norm optimisation approach to SDGs: tackling poverty by acting on discrimination

Summary of the chapter

Policies that seek to mitigate poverty by acting on equal opportunity have been found to aggravate discrimination against the poor (aporophobia), since individuals are made responsible for not progressing in the social hierarchy. Only a minority of the poor benefit from meritocracy in this era of growing inequality, generating resentment among those who seek to escape their needy situations by trying to climb up the ladder. Through the formulation and development of an agent-based social simulation, this study aims to analyse the role of norms implementing equal opportunity and social solidarity principles as enhancers or mitigators of aporophobia, as well as the threshold of aporophobia that would facilitate the success of poverty-reduction policies. The ultimate goal of the social simulation is to extract insights that could help inform and guide a new generation of policy making for poverty reduction by acting on the discrimination against the poor, in line with the UN "Leave No One Behind" principle. An "aporophobia-meter" will be developed and guidelines will be drafted based on both the simulation results and a review of poverty reduction policies at regional levels.

6.1. Introduction

This chapter of the thesis is a proposal of future research, which has been submitted to the International Joint Conference of Artificial Intelligence (IJCAI). Through the period of my PhD I have worked in different projects, creating as a result an interdisciplinary network of researchers on the topic of AI ethics and bias, with focus on aporophobia. This particular work constitutes an example of one of the lines of research that have been created in the framework this thesis in a collaborative way with other researchers. The coauthors of this research proposal are Nieves

Montes, Carles Sierra and Nardine Osman (from the Institut d'Investigació en Intel·ligència Artificial, IIIA-CSIC) and Flavio Comim (Universitat Ramon Llull, IQS School of Management).

6.1. Problem statement

Traditional poverty reduction policies have proved ineffective in the last decades. Despite the enormous efforts devoted to the redistribution of wealth, 700 million people, or 10% of the global population, still live in extreme poverty and evidence suggests that global poverty could increase by as much as 8% as a result of COVID-19 crisis, according to the United Nations (UN). The principles of equal opportunity and solidarity are key pillars of our welfare states and have been the main political answer to reduce poverty in terms of distributive justice. However, the rhetoric of equal opportunity has also been associated with the stigmatisation of the poor and the uneducated, who are considered blameful for not climbing up the social ladder [Sandel, 2020].

In an era of increasing inequality, the 1% top incomes saw a growth between 80% and 240% from 1980 to 2018 (Piketty 2020). Meanwhile, the poor are told that they have the opportunity to prosper if they study at a good university and work hard (Mounk 2017). However, meritocracy has not worked as expected. Only 7% of United States citizens from the 20% lower rents reach the 20% top rents within their lifetimes (Chetty et al., 2014). Results are not better in many European countries (Germany presents even lower social mobility than the US (OECD 2018). The difference, however, is that Europeans tend to underestimate social mobility whereas in the US social mobility is overestimated (Alesina et al. 2018).

It needs to be clarified that, even if policy makers tried their best to create an atmosphere close to perfect equal opportunity, there is no such thing since in practice, from the moment of birth, individuals are exposed to different environments (Fishkin, 2014). Furthermore, the rhetoric of equal opportunity can even constitute an obstacle to pass and implement policies aimed at reducing poverty for the so-called "undeserving poor" (Everatt, 2008), forcing policy makers to

determine which poor are victim of the circumstances ("luck egalitarism"), and therefore deserving aid, and which are responsible for their poverty (Anderson, 1999). As an undesired effect, a social stigma associated with the poor, aporophobia (Cortina 2017), is aggravated by blame. This has important psychological consequences for the well-being of people in need, who are avoided, discriminated or even attacked as a result of this socially-shared prejudice which leads to resentment, insecurity, self-hate.

The study of discrimination against the poor and its effect on poverty reduction has not received the attention it deserves in the literature. The Spanish philosopher Adela Cortina coined the term aporophobia to describe the rejection of the poor (Cortina 2017), and the concept was included in the Spanish legal framework as an aggravating factor for hate crimes in 2021 (Bolet´ın Oficial del Estado, 2021). As an example of the ubiquity of discrimination against the poor that concerns the AI community, chapter 4 of this thesis presents a pioneer study to provide evidence of bias against the poor in Word2Vec and GloVe embeddings by using word vector representations. However, studies are still required to provide evidence on whether aporophobia hinders the success of poverty reduction policies and can be considered an obstacle for the achievement of the first UN Sustainable Development Goal (SDG), poverty eradication.

Since testing the effectiveness of changes in policy and the resulting outcomes in real-life scenarios is an unfeasible process with great social and ethical repercussions, social simulation models emerge as powerful tools that, if well-formulated, can help assess the impact that new regulations have on a community. A recent example on the use of agent-based models to inform policy making concerns the COVID-19 pandemic and the simulation of non-pharmaceutical interventions on the evolution of infections (Hinch et al. 2021).

In this context, this multidisciplinary project aims to answer the following question: (1) what is the impact of prescriptive norms (aka policy measures) related to equal opportunity and social solidarity on aporophobia at a macro level?; and (2) to what extent does aporophobia at a micro

level influence the effectiveness of poverty reduction policies?

We will study this problem in the context of parametric norms, where solutions for optimal levels of equal opportunity and redistribution of resources can possibly be obtained using search and optimisation techniques, such as meta-heuristics. Previous work on agent-based modelling for the evaluation and optimisation of normative systems of income transfer (i.e. tax collecting and redistribution) exists in the literature (Sallila 2010). However, it tackles poverty reduction from a merely resource redistribution perspective. As detailed in our introduction, the resentment created by such measures often aggravates the discrimination against the very people whom these policies are supposed to help, unfortunately rendering them ineffective.

The main innovation of our project, then, consists in the introduction of aporophobia both as a macro indicator that the norms in place ought to minimise, and as an individual attitude towards the acceptance of such norms. Prior to that, we will develop a conceptual framework and gather empirical data on aporophobia, based on the state of the art studies on prejudice, discrimination and bias (Pettigrew, 2021; Paolini et al., 2021). We expect that this work will be a key addition to traditional poverty reduction models.

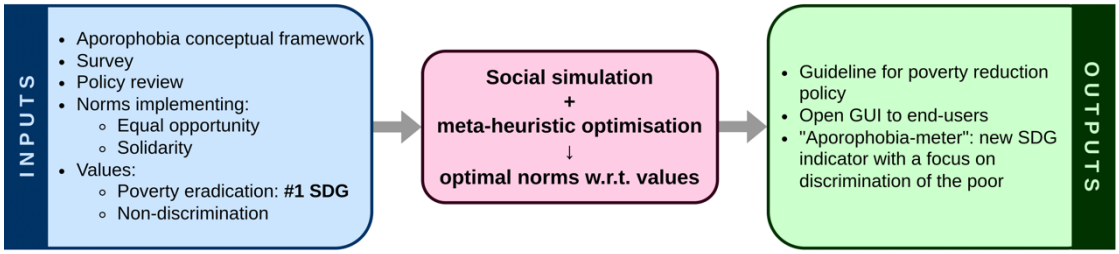

Figure 1: a norm optimisation approach to SDGs. Tackling poverty by acting on discrimination

6.2. Societal benefits and target SDGs

Assessing the impact of aporophobia on poverty levels opens a completely new path to tackle poverty reduction, from the point of view of non-discrimination measures and public awareness.

Poverty could therefore be mitigated not only by aiding the poor through redistributive justice, but also by mitigating the existing discrimination.

Provided we succeed in modelling the relationship between equal opportunity, aporophobia and poverty, social simulation models will allow to investigate the role of discrimination in poverty reduction. We will define guidelines and indicators to mitigate poverty that will possibly include tackling aporophobia, in the case that aporophobia is found to be a main barrier towards poverty reduction. The resulting societal benefits can be grouped into the following categories:

- Poverty reduction – number 1 SDG: We expect to provide evidence that aporophobia is a brake for poverty reduction policies. When the poor are considered responsible for their fate, poverty reduction policies may be not be well received by the public, compromising their effectiveness. The work developed in this project will seek to quantify the effectiveness of poverty reduction policies as a function of the aporophobic levels of the population and provide recommendations for policy-making taking these results into account.
- Reduce resentment: An important characteristic of poverty goes beyond the lack of material goods. As Nobel Laureate Amartya Sen states, a person is not poor when he or she can carry out a meaningful life with dignity (2001). However, the rhetoric of equal opportunity in a time of growing inequality can lead to the frustration of some of the population in need (Hochschild 2018). As a result of the guidelines and indicators provided as outputs of this study, awareness campaigns can be put in place to mitigate the social stigma associated to the poor. This constitutes a new paradigm on poverty reduction approaches, since the focus is not only on the poor, but also on the non-poor and society in general.
- Enhanced sense of community: This study will provide indication for the optimal equilibrium between norms implementing equal opportunity and social solidarity. Having equal opportunity as the main answer to poverty creates a high level of competition

among citizens, (Mounk 2017; Sandel 2020b), enhanced by increasing inequality in the era of globalisation [Osnos 2014]. When other non- competitive solutions are in place in terms of contributive justice, such as recognising the value of jobs independently of the salaries, individuals perceive social recognition and part of a community (Smith and Deranty 2011).

6.3. Goals and methods

The goals that this collaboration aims to achieve involve two distinct but complementary disciplines: the social sciences (in particular philosophy, psychology and welfare economics), and computer science and AI.

Methodologically, we intend to follow the framework originally presented by the authors in (Montes and Sierra, 2021; Montes and Sierra, 2022) to develop such a social model. In general terms, the method consists of the following steps. First, the state features that are relevant in the domain where the simulation is focused are to be defined. Provided that we are interested in poverty from a multidimensional perspective, potential features will look not only for resource deprivation, but also health and education failures, among others. These definitions will require a close collaboration between the AI and the the social sciences team members.

Second, the prescriptive norms (aka the regulations) that govern the society of agents have to be defined as well as the optimisable parameters they are tied to, i.e. the normative parameters. The union of all the normative parameters, together with their bounds and constraints, provides the space where a search algorithm will look for the optimal levels. In this project, we intend to focus on norms implementing social solidarity and equal opportunity. Therefore, it will be necessary to formulate how these norms operate and what is their impact on the behaviour of the agents.

Third, the semantics of the values of interest have to be defined. We understand values as abstract general concepts, whose meaning is grounded in a particular context (a socio-economic

simulation in this case) by a particular goal or function. Values, hence, are operationalised through these goals and functions, which serve to evaluate the states of the system and the achieved outcomes. The semantics of such values will provide us with the optimisation target as the alignment of a candidate normative system with respect to the value of interest. For this project, we are interested in values related to the eradication of poverty and the reduction of aporophobia. For example, a wide variety of indicators for economic equality are available (Cowell 2011), of which the authors will need to discern and pick the most suitable ones or possibly come up with their own indicators with a stronger focus on discrimination.

Fourth and final, a suitable AI technique has to be chosen, considering the scale and computational requirements of the resulting social model, to automatically perform the search over the space of normative parameters with the target on aporophobia and poverty reduction. We will be looking for the optimal levels of solidarity and equal opportinity that would maximise the promotion of the values of focus, i.e. “minimum poverty” and “maximum fairness (understood an idea desired value of non-discrimination against the poor)”. In order to have a meaningful assessment of the effect of a candidate normative system, a sampling technique over several runs of the model, such as Monte Carlo sampling, is necessary, regardless of the search algorithm of choice. In the past, the authors have settled for a Genetic Algorithm (GA) to conduct the search (Montes and Sierra 2021). Although we do not discard, right off the bat, other optimisation techniques, in particular meta- heuristics ones such as Simulated Annealing, many of the reasons that made us settle for GAs in our previous work are applicable in this project, namely their versatility for optimisation over continuous, discrete and hybrid (both continuous and discrete variables) domains [Luke, 2013]. It is worth noting that GAs are also very suited for parallelisation, as the recombination of parent normative systems for the creation of the new generation can be easily distributed over several computing nodes. This is a clear point in favour of sticking to this class of search algorithms, as we foresee that the scale of a meaningful model, that is able to reflect with some degree of fidelity real-life scenarios, will require intensive

computational resources.

Additionally, a major extension to the norm optimisation framework in (Montes and Sierra 2021) needs to be developed in order to account for the aporophobic attitudes of the agents in the system. As a preliminary idea, we contemplate modelling aporophobia at the micro level as an internal psychological agent construct linked to the ranges of the normative parameters an individual is willing to accept. This, in turn, conditions their willingness to abide by the established norms. We would like to emphasise, however, that the computational construct to introduce aporophobia into the model will be reviewed and refined after the authors have finalised the conceptual framework of aporophobia and obtained empirical evidence of the nature of the phenomenon through surveys. We expect that this prior revision will make the agent-based model more sound and better grounded in the state of the art in social psychology.

The final objective that the collaboration pursues is the closure of the feedback from the computational results to the social sciences. The ultimate goal of the social simulation is to extract insights that could help inform and guide a new generation of policies for poverty reduction. To achieve this goal, a policy review of regional legislation (including a comparative analysis of the Global North and the Global South) is to be conducted. Comments, suggestions and guidelines will be drafted based on the simulation results. If, hypothetically, the simulation results conclude that no level of aporophobia should be allowed in order for poverty-reduction measures to be effective, recommendations would be made on targeting the aporophobia levels of the population through communication campaigns. Other insights from the simulation results would need to be translated back into the realm of the social sciences in terms of other measures, possibly targeting the adequate balance between social solidarity and equal opportunity.

As an final touch to the project, and with the intention to reach as many policy makers and social science scholars as possible, we will develop a graphical user interface (GUI) to the social simulation or a lightweight version of it. With a GUI, anyone could interactively switch parameter

values or define what values are they interested in promoting and examine the effect that those changes would have on a community of agents. Such a step would make the technical aspects of this research much more accessible to the people that should extract insights and act upon it.

6.4. Challenges and risks

We identify five main risks to be tackled during the execution of the proposal, which we itemise alongside with their mitigating measures in Table 1. These can be categorised into two classes. The first concerns the quality of the agent-based model (rows one, two and three in Table 1). In order to ensure that the formulated model is sound, unbiased and produces relevant insights, it will be grounded on the state-of-the-art on discrimination, poverty, equal opportunity and social solidarity literature, as well as the incipient literature on aporophobia, such as chapters 2 and 3 of this thesis. In the context of this collaborative project, the authors will also develop their own survey to obtain empirical data on the levels of aporophobia in human subjects. However, basing the agent-based model on data extracted from a small subset of the world population (namely people from western countries or the Global North) would render the model biased towards the necessities of these societies and the policy recommendation derived from the results irrelevant to a wide range of audiences. To avoid this, the model needs to be adaptable to local characteristics which, in turn, will require conducting the survey on subjects from many different backgrounds and developing the simulation in a modular way.

Another challenging task related to the formulation of the model is the definition of the semantics of values minimum poverty and maximum fairness (non-aporophobia), which we want to embed into the system. This is a very important point since it defines the objectives that we would like to achieve in the simulations and will direct the recommendations derived from that. Relative and absolute concepts of poverty will be used, in line with the state-of-the-art literature on human development, considering also the impact of inequality. The analysis of poverty will be multidimensional and based on a lack of basic capabilities, as described by Nobel Laureate

Amartya Sen, to avoid generating recommendations only based on reaching above poverty threshold incomes, which could have perverse effects.

The second category of risks to be tackled concerns the availability of resources to carry out the project, which are being mitigated as described in Table 1.

6.5. Evaluation criteria

Once the agent-based model is implemented and the methodology is applied, a set of optimal norms will be obtained. These norms will implement a poverty reduction policy strategy and their effects will be quantified. In addition to these results, the authors have also developed a set of evaluation analytical tools to examine in depth the resulting normative systems, optimised for poverty reduction.

The first of these evaluation tools consists in conducting a Shapley value analysis of the resulting optimal norms. This approach is grounded on the very well established field of co-operative game theory, and considers that every individual norm (aka social solidarity and equal opportunity) is a member of a coalition, i.e. the normative system at large. Then, it is possible to apply the definition of Shapley value, taking as the worth of coalitions the alignment that these normative systems are able to achieve with respect to poverty reduction targets. Such a computation yields a quantitative evaluation on the importance of every individual norm when it comes to poverty reduction. This is a very informative metric that helps discern the mechanisms by which the optimised normative systems are achieving their targets.

The second of these evaluation tools regards the compatibility of poverty reduction with other values that policy-makers and scholars might deem relevant, such as social mobility. In our previous research, we have encountered that normative systems that are highly optimised for some value can be very oblivious for others (Montes and Sierra 2022). Additionally, if the users of the model deem it interesting, it is also possible to perform optimisation searches with the target being

not a particular value such as poverty reduction, but the compatibility degree among several values. This type of search will produce the normative system that is the best compromise for a set of different values.

However, this computational evaluation is not the whole story, as we will need to check if humans accept the norms that have been obtained, no matter what is their initial level of aporophobia and whether, after analysing them, their levels of aporophobia get reduced. In order to perform this evaluation, surveys will be constructed to measure and follow up on the levels of aporophobia, before and after exposing the subjects to the reading and analysis of the norms. A well-known problem of questionnaires that seek to measure ethical topics in the population is bias, since individuals do not tend to be completely honest. This will been taken into account in the questionnaire design according to the state-of-the-art on survey design on moral topics. [Greenwald et al 2009]. It is imaginable that if the results leave room for improvement, further iterations of modelling, optimisation, and evaluation would be run.

Table 1: Potential risks that will encountered during the execution of the process and the corresponding mitigating actions.

Risk description	Mitigating actions
Ungrounded social model	The model will be based on the state-of-the-art on discrimination and poverty, as well as empirical data on aporophobia obtained from the author's own survey. Also, the model will be based on a methodology that has been proven successful in the past (Montes and Sierra, 2021; Montes and Sierra, 2022).
Biased social model	The survey to obtain data will be used at a global level and, when used regionally, the resulting recommendations will be tailor-made by adapting the model to local characteristics. Demographic data will be collected through the survey to ensure the representativity of in- come levels, education and

	professional backgrounds. Especial attention will be paid to the representation of historically discriminated groups.
Irrelevant results	A review of the poverty reduction policy framework at a regional and global level (including the Global North and the Global South) will be performed so that the guidelines resulting from the project provide added value to international and regional NGOs and government officials.
Lack of financial resources	The two direct costs of the project are human resources and computational equipment. These are available to the existing team since this project has been incorporated as an internal objective.
Lack of computating power	The IIIA team has the support of an HPC service and access to the Ars Magna cluster. In order to leverage these computational resources, a search strategy amenable to parallelisation, such as the Genetic Algorithms discussed in Section 3, will be implemented.

6.7. Long term impact of the SDGs

This line of research aims at opening a completely new path to tackle poverty reduction at a macro-global, meso-national and micro-individual levels, which is UN's #1 SDG (poverty eradication) and is clearly related to the other 16 SDGs, due to the multidimensional nature of poverty: #2 SDG (zero hunger), #8 SDG (decent work and economic growth) and #10 SDG (reduced inequalities). The proposed work informs about the optimal levels of norms related to "equal opportunity" and "social solidarity" to attain the values of "minimum poverty" and "maximum fairness" (understood within the framework of the project as an ideal desired value of non- aporophobia). The research line also allows to explore alternative approaches to poverty reduction within an original AI simulation context, based on existing data and providing

recommendations that aim to be potentially applied in a real-life scenarios, both regionally and globally.

It has been suggested that the discrimination against the poor (aporophobia) could have an impact on poverty at different levels. From a macro-international perspective, the developing countries are considered responsible for their fate, instead of working towards a global equilibrium in areas such as international commerce, cooperation among countries and financial markets. At a meso-national level, aporophobia hinders the effective implementation of poverty reduction measures. Finally, at a micro-personal level, the self-depreciation of the poor is an additional obstacle to improve their eco- nomic situation. By providing evidence that aporophobia constitutes an obstacle for poverty reduction, the study opens the opportunity to a completely new set of measures to reduce poverty, acting on the socially shared prejudices towards the poor. The focus of the problem (and the solution) would be not only on the poor, but also on the non-poor and the society as a whole.

The "aporophobia-meter" will allow to measure and follow up the evolution of aporophobia, enabling policy makers to relate it with poverty levels throughout time. A long-term goal of the study is to encourage a virtuous circle where less discrimination against the poor lead to a higher effectiveness of poverty reduction policies at global and regional levels.

Chapter 7. Conclusions and future research lines

This thesis offers a conceptual framework of the circumstances that explain aporophobia and describes how aporophobia is a discriminatory phenomenon aggravated as a result of industrial and AI capitalism social recognition orders. Whereas in feudal societies social status was determined at birth, in industrial capitalism status is defined according to the achievement principle, publically represented by wealth and credentialism. Individuals, therefore, are to some extent made responsible for being poor. The feelings of shame and self-depreciation are even more exacerbated as a result of the rhetoric of equal opportunity which is an essential part of welfare states.

This study also offers the first empirical evidence of the existence of aporophobia in a social-networks-spontaneous scenario and in the pretrained embeddings of Google, Twitter and the Wikipedia. This is particularly relevant since these embeddings are used to develop apps in critical areas such as health, justice or education around the world. Further studies need to be performed particularly to measure and follow up the phenomenon in different geographical regions, cultural contexts and AI systems, especially analysing the relationship between bias against the poor, poverty and inequality indicators. Additionally, studies need to be carried out to analyse the phenomenon of plutofilia, since according to Allport (1954) overestimation of what we love occurs before the underestimation of the contraries.

While it is necessary to offer a critical analysis of the current scenario in terms of AI and bias against the poor, this thesis also aims to shed some light on specific ways forward. Therefore, a pro-ethical process to identify and mitigate bias within the development of AI systems is defined, which intends to provide support to development teams and seek the involvement of stakeholders, including historically discriminated groups. While AI has traditionally been considered an

artifact, there is a wide field of research to develop on how it relates to human experience, for which technology can be dealt with as a process where all stakeholders should participate and create a narrative, beyond the technical description. In this sense, a future line of research will be to test the proposed Artificial Intelligence Bias Mitigating Process (AIBMP) in a real-context scenario to find out whether the perception of fairness and trust among users increases.

The thesis describes how AI systems are supporting and often exacerbating the existing bias against the poor. In addition to the digital divide, AI capitalism often degrades the working conditions of the most vulnerable and it has been described how it enhances biases against the poor "by design" even for essential social services. However, dealing with AI bias is not only a technological issue since, to start with, there is not a universal perception of fairness, which is dependent on the context and the individual. Bias in AI systems is only the tip of an iceberg showing the prejudices that are culture-dependent, for which this work offers a multidisciplinary analysis of AI bias informed on political philosophy, psychology, social economics, sociology, ethics and business analysis. From this conceptual analysis, proposals to tackle the phenomenon of bias and AI, with specific focus on aporophobia, are suggested, aimed at acting on the structure of the AI business, supporting social activism and cooperative entrepreneurship or modifying the value chain, such as encouraging the creation of intermediaries that manage the use of data according to users' instructions, as suggested by the World Economic Forum (2022). As it happens with AI systems, businesses are often considered "amoral", but they obviously also convey values. Therefore, an additional future line of research is the study of the potential diversification of AI platforms and the specific ethically tagged content to openly communicate decisions and trade-offs dealing with fairness and bias in order to cater for a diversity of value-oriented users, including positive discrimination of historically discriminated groups and the representation of the Global South. While AI for Good aims at contributing to the achievement of the Sustainable Development Goals, this line of research would have user's and relational empowerment as a goal. We have named this line of research "People's AI".

The lack of neutrality of AI systems is also analysed in this study. AI responds to historical values and generates new value recognition orders based on the capacity to create information flows through which data commodities can reach a market. It also influences the political context and generates economic activity. In fact, AI systems incorporate morality, since they provide considerations about what is good and what is a duty (Ausín and Robles Carrillo 2021), they transform people's habits and the supreme value they convey is the information flow. AI is changing the players and the rules of the game in the so-called "surveillance capitalism" (Zuboff 2019), since data based on the users' behaviour is appropriated for free to generate profit. However, current AI is far from the human capacity of reasoning. One could say that it has abilities without real capacity to understand and, especially, without common sense (Cortina 2019 ; Mántaras 2020). However, algorithmic decision making is being increasingly used in critical fields such as health, education and justice in a context of technology overestimation and lack of critical questioning. This is often promoted by private companies managing the almost totality global data in a market characterised by fierce competition and oligopoly, in what has been called "economic platformisation". Another future line of research, therefore, will be to study AI not under an instrumental view (that is as a means to reach one's aims) but as a relational moral concern in itself. AI needs to be understood within a socio-technical eco-system (Ausín 2021), where human beings relate to other human beings, to technological devices, algorithms and data. In this new context one should aim at enhancing digital agency and capabilities, understood under the sense defined by Sen (2001) and Nussbaum (2012). However, capabilities in the online are currently limited due to the lack of transparency in what has been described as the AI black-box and the often unnoticed use of personal data to influence individual behaviour (Zuboff 2019).

Finally, AI constitutes an opportunity not only to measure and follow up specific types of bias through social networks and pre-trained embeddings (NLP) but to actually offer alternative ways to act towards the UN SDGs. From the literature analysis, this thesis explains that aporophobia constitutes an obstacle to reduce poverty at macro-international, national and micro-personal

levels. Poverty reduction policies have been documented to be more restrictive as a result of aporophobia and the blamefulness of the poor. Politicians need to somehow justify which poor are deserving and not deserving aid and this phenomenon adds up to the burden carried by the historically discriminated groups in terms of gender, ethnicity or sexual orientation. The last line of research derived from this thesis aims to provide empirical evidence that aporophobia constitutes an obstacle for poverty reduction through an AI norm optimisation approach. While traditional redistributive poverty reduction policies have proved ineffective in the last decades, acting on discrimination against the poor (by using AI-supported-decision-models) can constitute a completely new path to work towards the first UN SDG (poverty eradication).

References

Adamuthe AC (2020) Improved Text Classification using Long Short-Term Memory and Word Embedding Technique. Int J Hybrid Inf Technol 13

Aggarwal N (2020) The Norms of Algorithmic Credit Scoring. SSRN Electron J. https://doi.org/10.2139/SSRN.3569083

Aghion PE, Caroli, Garcia-Penalosa C (1999) Redistribuion inequality and growth. J Econ Lit 37:1615–1660

Alesina A, Stantcheva S, Teso E (2018) Intergenerational Mobility and Preferences for Redistribution. Am Econ Rev 108:521–54. https://doi.org/10.1257/AER.20162015

Alessina A, Glaeser EL (2013) Fighting Poverty in the US and Europe. Oxford University Press, Oxford

Algorithm W (2021) AI Ethics Guidelines Global Inventory. In: Algorithm Watch. https://inventory.algorithmwatch.org/. Accessed 4 Dec 2021

Allison G, Schmidt E (2020) Is China Beating the U.S. to AI Supremacy?. Belfer Center for Science and International Affairs.

Allport GW (1954) The nature of prejudice. Basic Books

Almeida F, Xexéo G (2019) Word Embeddings: A Survey. https://arxiv.org/abs/1901.09069v1

Ananny M, Crawford K (2018) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. undefined 20:973–989. https://doi.org/10.1177/1461444816676645

Anderson ES (1999) What is the point of equality? Ethics 109:287–337.

https://doi.org/10.1086/233897/0

Anshari M, Almunawar MN, Masri M, Hrdy M (2021) Financial Technology with AI-Enabled and Ethical Challenges. Society 58:189–195. https://doi.org/10.1007/S12115-021-00592-W

Applebaum LD (2001) The Influence of Perceived Deservingness on Policy Decisions Regarding Aid to the Poor. Polit Psychol 2

Aristóteles (2012) Política. Espasa Libros, Barcelona

Arneson RJ (1997) Egalitarianism and the Undeserving Poor. J Polit Philos 5:327–350

Arnold T, Scheutz M (2018) The "big red button" is too late: an alternative model for the ethical evaluation of AI systems. Ethics Inf Technol 20:59–69. https://doi.org/10.1007/S10676-018-9447-7

Aumer-Ryan K, Hatfield E (2007) The design of Everyday Hate: A Qualitative and Quantitative Analysis. Interpersona1

Ausín T, Robles Carrillo M (2021) ÉTICA Y DERECHO EN LA REVOLUCIÓN DIGITAL. Rev Diecisiete Investig Interdiscip para los Objet Desarro Sostenible 04:15–28. https://doi.org/10.36852/2695-4427_2021_04.00

Ausín T, (2021) ¿Por qué ética para la Inteligencia Artificial? Lo viejo, lo nuevo y lo espúrio. Ediciones Universidad Valladolid.

Azmanova A (2020) Capitalism on edge : how fighting precarity can achieve radical change without crisis or utopia. Columbia University Press

Barocas S, Selbst AD (2016) Big Data's Disparate Impact. SSRN Electron J. https://doi.org/10.2139/SSRN.2477899

Becker GS (1965) A Theory of the Allocation of Time. Econ J 75:493–517

Benjamin R (2019) Captivating technology: Race, Carceral Technocience, and Liberatory

Immagination in Everyday Live. Duke University Press, Durham

Beukeboom CJ, Burgers C (2019) How stereotypes are shared through language: A review and introduction of the Social Categories and Stereotypes Communication (SCSC) framework. Rev Commun Res 7:1–37. https://doi.org/10.12840/ISSN.2255-4165.017

Binns R Algorithmic Accountability and Public Reason. https://doi.org/10.1007/s13347-017-0263-5

Blodgett SL, Barocas S, III HD, Wallach H (2020) Language (Technology) is Power: A Critical Survey of "Bias" in NLP. 5454–5476. https://doi.org/10.18653/V1/2020.ACL-MAIN.485

Bojanowski P, Grave E, Joulin A, Mikolov T (2016) Enriching Word Vectors with Subword Information. Trans Assoc Comput Linguist 5:135–146. https://doi.org/10.1162/tacl_a_00051

Boletín Oficial del Estado. Circular 7/2019, de 14 de mayo, de la fiscalía general del estado, sobre pautas para interpretar los delitos de odio tipificados en el artículo 510 del código penal. BOE-A-2019-7771, 2021. Ministerio de la Presidencia, Relaciones con la Cortes y Memoria Democrática (Gobierno de España).

Bolukbasi T, Chang K-W, Saligrama V, et al (2016) Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. ArXiv ID: 1607.06520v1

Bourdieu P (2010) Distinction : a social critique of the judgement of taste / Pierre Bourdieu ; translated by Richard Nice, with a new introduction by Tony Bennett. Routledge Classics

Boyd R, Richerson PJ (2009) Culture and the evolution of human cooperation. Philos Trans R Soc B Biol Sci 364:3281–3288

Bradford A (2020) The Brussels effect: How the European Union rules the world. Oxford University Press

Burkett I (2000) Beyond the "information rich and poor": futures understandings of inequality

in globalising informational economies. Futures 32:679–694

By A, Silberg J, Manyika J (2019) Notes from the AI frontier: Tackling bias in AI (and in humans). Mckinsey Global Institute

Caliskan A, Bryson JJ, Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. Science (80-) 356:183–186. https://doi.org/10.1126/science.aal4230

Camacho-Collados J, Pilehvar MT (2020) Embeddings in Natural Language Processing. 10–15. https://doi.org/10.18653/V1/2020.COLING-TUTORIALS.2

Cameron LD, Rahman H (2021) Expanding the Locus of Resistance: Understanding the Co-constitution of Control and Resistance in the Gig Economy. https://doi.org/101287/orsc20211557 33:38–58. https://doi.org/10.1287/ORSC.2021.1557

Card D, Smith NA (2020) On Consequentialism and Fairness. Front Artif Intell 3:34. https://doi.org/10.3389/FRAI.2020.00034/BIBTEX

Carman M, Rosman B (2021) Applying a principle of explicability to AI research in Africa: should we do it? Ethics Inf Technol 23:107–117. https://doi.org/10.1007/S10676-020-09534-2

Chancel L, Piketty T (2021) GLOBAL INCOME INEQUALITY, 1820-2020: THE PERSISTENCE AND MUTATION OF EXTREME INEQUALITY. J Eur Econ Assoc. https://doi.org/10.1093/jeea/jvab047

Charles Lewis Taylor, Hudson MC (1972) World Handbook of Political and Social Indicators II. World Handbook of Political and Social Indicators Series

Chatila R, Havens JC (2019) The IEEE global initiative on ethics of autonomous and intelligent systems. Intell Syst Control Autom Sci Eng 95:11–16. https://doi.org/10.1007/978-3-030-12524-0_2

Chetty R, Hendren N, Kline P, et al (2014) Where is the land of Opportunity? The Geography of Intergenerational Mobility in the United States. Q J Econ 129:1553–1623. https://doi.org/10.1093/QJE/QJU022

Chiappa S, Jiang R, Stepleton T, et al (2020) A General Approach to Fairness with Optimal Transport. Proc AAAI Conf Artif Intell 34:3633–3640. https://doi.org/10.1609/AAAI.V34I04.5771

Coeckelbergh M (2022) The Political Philosophy of AI. Polity

Coeckelbergh M (2021) Time Machines: Artificial Intelligence, Process, and Narrative. Philos Technol 34:1623–1638. https://doi.org/10.1007/S13347-021-00479-Y

Comim F, Borsi MT, Valerio Medoza O (2020) The Multi-dimensions of Aporophobia.MPRA

Cortina A (2007) Etica de la razón cordial: educar en la ciudadanía en el siglo XXI. Ediciones Nobel (Kindle)

Cortina A (2017) Aporofobia, el rechazo al pobre. PAIDOS, Barcelona

Cortina A (1986) Ética mínima : introducción a la filosofía práctica. Tecnos, Madrid

Cortina A (2021) Ética cosmopolita: Una apuesta por la cordura en tiempos de pandemia (Estado y Sociedad). Paidós

Cortina A (2019) Ética de la Inteligencia Artificial. Anales de la Real Academia de Ciencias Morales y Políticas 96 : 24.

Cortina A (1991) La moral del camaleon: ética para nuestro fin de siglo. Espasa Calpe, Madrid

Couldry N, Mejías UA (2019) The Costs of Connection: How Data is Colonizing Human Live and Appropriating it for Capitalism. Standford University Press, Standford

Cowell F (2011). Measuring Inequality. Oxford University Press.

Cowls J, Tsamados A, Taddeo M, Floridi L (2021) The AI Gambit — Leveraging Artificial Intelligence to Combat Climate Change: Opportunities, Challenges, and

Recommendations. SSRN Electron J. https://doi.org/10.2139/SSRN.3804983

Crenshaw K (1991) Stanford Law Review Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of. Source Stanford Law Rev 43:1241–1299

D. Grusky (ed) (2014) Social Stratification. Westview Press

Dabla-Norris E, Kochhar K, Suphaphiphat N, et al (2015) Causes and consequences of income inequality: a global perspective. International Monetary Fund

Darwin C (2004) The Descent of Man and Selection in Relation to Sex. Pengin Classics.

Davis JL, Williams A, Yang MW (2021) Algorithmic reparation. Big Data Soc 8:. https://doi.org/10.1177/20539517211044808

De Kloet J;, Poell T;, Zeng G;, Chow YF (2019) The platformization of Chinese Society: infrastructure, governance, and practice. Chinese J Commun. https://doi.org/10.1080/17544750.2019.1644008

De Mauro A, Greco M, Grimaldi M (2016) A formal definition of Big Data based on its essential features. Libr Rev 65:122–135. https://doi.org/10.1108/LR-06-2015-0061

Derrida J (1978) Writing and difference. Univ Chicago Press

Dignum V (2022) Relational Artificial Intelligence. ArXiv ID: 2202.07446v1

Dijck J van, Poell T, Waal M de (2018) The platform society : public values in a connective world. Oxford University Press

Durkheim É (1960) The division of labour in society. Free Press of Glencoe, Illinois

Dwork C, Hardt M, Pitassi T, et al (2011) Fairness Through Awareness.ArXiv: 1104.393

Dworkin R (1981) What is equality? Part 2: Equality of resources. Philos Public Aff 10

Dyer-Witheford N (2019) Inhuman Power: Artificial Intelligence and the Future of Capitalism. Pluto Press, London

Eberhardt J (2020) Biased. Penguin Books

Echevarría B (2011) Modernidad y Blanquitud. Biblioteca Era, Ciudad de México

Edelman Trust Barometer (2020) Special report: trust in technology

Esquembre CO (2019) La aporofobia como desafío antropológico. De la lógica de la cooperación a la lógica del reconocimiento. Daimon Rev Int Filos 215–224

Ess C (2020) Digital media ethics. Wiley

Eubanks V (2018) Automating inequality. How high-tech tools profile, police and punish the poor. St. Martin's Press

European Comission (2019) Factsheet on Artificial Intelligence - Artificial Intelligence for Europe

European Commission (2020) White paper on artificial intelligence: a European approach to excellence and trust. White Paper COM(2020) 65 final, European Commission, Brussels, February 2020.

European Commission (2021) Artificial Intelligence Act. "Proposal for a regulation of the European Parliament and the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts."

Everatt D (2009) The Undeserving Poor: Poverty and the Politics of Service Delivery in the Poorest Nodes of South Africa. Politikon 35:293–319

Fazelpour S, Lipton ZC (2020) Algorithmic fairness from a non-ideal perspective. AIES 2020 - Proc AAAI/ACM Conf AI, Ethics, Soc 57–63. https://doi.org/10.1145/3375627.3375828

Fishkin J (2014) Bottlenecks. Bottlenecks. https://doi.org/10.1093/ACPROF:OSO/9780199812141.001.0001

Floridi L (2015) The onlife manifesto: Being human in a hyperconnected era. Onlife Manif Being Hum a Hyperconnected Era 1–264. https://doi.org/10.1007/978-3-319-04093-6

Floridi L (2019a) Translating Principles into Practices of Digital Ethics: Five Risks of Being

Unethical. Philos Technol 2019 322 32:185–193. https://doi.org/10.1007/S13347-019-00354-X

Floridi L (2019b) Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. Philos Technol 2019 322 32:185–193. https://doi.org/10.1007/S13347-019-00354-X

Floridi L (2021) The European Legislation on AI: A Brief Analysis of its Philosophical Approach. SSRN Electron J. https://doi.org/10.2139/SSRN.3873273

Floridi L (2019c) The logic of information. Oxford University Press, Oxford

Floridi L, Cowls J, Beltrametti M, et al (2018) AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds Mach 28:689–707. https://doi.org/10.1007/s11023-018-9482-5

Floridi L, Taddeo M (2016) What is data ethics? Philos Trans R Soc A Math Phys Eng Sci 374:. https://doi.org/10.1098/RSTA.2016.0360

Folbre N (2021) The Rise and Decline of Patriarchal Systems. An Intersectional Political Economy. Verso, London - New York

Fraser N, Honneth A (2003) Redistribution or recognition? A political -philosophical exchange. Verso Books

Freeman JB, Ambady N (2011) A Dynamic Interactive Theory of Person Construal. Psychol Rev 118:247–279. https://doi.org/10.1037/A0022327

Freeman RE, Martin K, Parmar B (2007) Stakeholder capitalism. J Bus Ethics 74:303–314. https://doi.org/10.1007/S10551-007-9517-Y

Fry H (2018) Hello world : being human in the age of algorithms. Penguin

Fuchs C (2018) Capitalism, Patriarchy, Slavery, and Racism in the Age of Digital Capitalism and Digital Labour. Crit Sociol 44:677–702. https://doi.org/10.1177/0896920517691108

Fuchs C (2020) Communication and Capitalism: A Critical Theory. Commun Capital A Crit Theory. https://doi.org/10.16997/BOOK45

Fukuyama F (2018) Against Identity Politics: The New Tribalism and the Crisis of Democracy. Foreign Aff 97:90–115

Fuster DGG, Brussel VU (2020) New EU financing instrument of up to €150 million to support European artificial intelligence companies | Shaping Europe's digital future. European Commission Press Release

Galor O, Moav O (2004) From physical to Human Capital Accumulation: Inequality and the Process of Development. Rev Econ Stud 71:1001-- 26

Gans HJ (1994) Positive functions of the underserving poor: uses of the underclass in America. Polit Soc 22:269–283

Gill I (2020) Whoever leads in artificial intelligence in 2030 will rule the world until 2100. Brookings

Goffman E (1963) Stigma Notes on the Management of Spoiled Identity. SIMON &amp SCHUSTER

Green B (2020) Algorithmic Realism: Expanding the Boundaries of Algorithmic Thought. Proc 2020 Conf Fairness, Accountability, Transpar. https://doi.org/10.1145/3351095

Green B, Chen Y (2019) Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. FAT* 2019 - Proc 2019 Conf Fairness, Accountability, Transpar 90–99. https://doi.org/10.1145/3287560.3287563

Green B, Hu L (2018) The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning. Mach Learn Debates Work 35th Int Conf Mach Learn

Greenwald, A.G, Poehlman T.A., Uhlmann E.L., and Mahzarin R. Banaji. (2009). Understanding and Using the Implicit Association Test: III. Meta-Analysis of Predictive Validity. Journal of Personality and Social Psychology, 97(1):17–41, jul 2009.

Gutiérrez L, Keith B (2019) A systematic literature review on word embeddings. Adv Intell Syst Comput 865:132–141. https://doi.org/10.1007/978-3-030-01171-0_12

Habermas J (1990) Moral consciousness and communicative action. Polity Press, London

Hacker P (2018) Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU Law (Common Market Law Review 2018, 1143-1185). Common Mark Law Rev

Hardt M, Price E, Srebro N (2016) Equality of Opportunity in Supervised Learning. ArXiv ID: 1610.02413v1

Harrison G, Hanson J, Jacinto C, et al (2020) An empirical study on the perceived fairness of realistic, imperfect machine learning models. FAT* 2020 - Proc 2020 Conf Fairness, Accountability, Transpar 392–402. https://doi.org/10.1145/3351095.3372831

Harvey D (2005) A Brief History of Neoliberalism. A Br Hist Neoliberalism. https://doi.org/10.1093/OSO/9780199283262.001.0001

Hauge M V., Stevenson MD, Rossmo DK, Le Comber SC (2016) Tagging Banksy: using geographic profiling to investigate a modern art mystery. J Spacial Sci 61:185–190. https://doi.org/10.1080/14498596.2016.1138246

Hegel GWF (1991) Elements of the Philosophy of Right. Oxford's Words Classics

Hinch R., Probert W.J.M., Nurtay A., Kendall M., Wyman C.t, Hall M., Lythgoe K., Bulas Cruz A., Zhao L., Stewart A., Ferretti L., Montero D., Warren J., Mather N., Abueg M., Wu N., Legat O., Bentley K., Mead T., VanVuuren K., Feldner-Busztin D., Ristori T., Finkelstein T.A., Bonsall D.G., Abeler-Dorner L, and Fraser C (2021). Openabm-covid19—an agent-based model for non-pharmaceutical interventions against covid-19 including contact

tracing. PLOS Computational Biology, 17(7):1–26, 07 2021.

HLEGAI (2019) High-Level Expert Group on Artificial Intelligence, EU - Ethics guidelines for trustworthy AI. Publications Office. High-Level Expert Group on AI - European Commission and Directorate General for Communications Networks, Content and Technology.

Hobbes T (2018) Leviatán. Mosaicum Books

Hochschild. A.R. (2018) Strangers In Their Own Land. Ingram Publisher Services, March 2018.

Hoffmann AL (2019) Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. https://doi.org/101080/1369118X20191573912 22:900–915. https://doi.org/10.1080/1369118X.2019.1573912

Honneth A (1996) The Struggle for Recognition. Polity Press

Ibáñez JC, Olmeda MV (2021) Operationalising AI ethics: how are companies bridging the gap between practice and principles? An exploratory study. AI Soc. https://doi.org/10.1007/S00146-021-01267-0

Ideo.org (2015) The field guide to human-centred design

Imbrie A, Kania E, Laskai L (2020) The Question of Comparative Advantage in Artificial Intelligence: Enduring Strengths and Emerging Challenges for the United States. https://doi.org/10.51593/20190047

International Telecmmunications Union (ITU) (2020) Measuring digital development. Facts and Figures. Geneva

Jiang L, Hwang JD, Bhagavatula C, et al (2021) Delphi: Towards Machine Ethics and Norms. ArXiv ID: 2110.07574

Jobin A, Ienca M, Vayena E (2019b) The global landscape of AI ethics guidelines. Nat Mach Intell 1:389–399. https://doi.org/10.1038/S42256-019-0088-2

Joseph K, Morgan JH (2020) When do Word Embeddings Accurately Reflect Surveys on our Beliefs About People? 4392–4415. https://doi.org/10.18653/v1/2020.acl-main.405

Kant I (2015) The Critique of Practical Reason. Cambridge University Press

Kellogg KC, Valentine MA, Ang` A, Christin A ALGORITHMS AT WORK: THE NEW CONTESTED TERRAIN OF CONTROL Work and Organization Studies MIT Sloan School of Management. Acad Manag Ann 2020:366–410. https://doi.org/10.5465/annals.2018.0174

Kelly-Lyth A, Stevens P (2022) Capital vs Labour? “Algorithmic Management” workers’ right, and the gig economy. Ethics for a Changing World Podcast.

Kiritchenko S, Zhu X, Mohammad SM (2014) Sentiment Analysis of Short Informal Texts. J Artif Intell Res 50:723–762. https://doi.org/10.1613/JAIR.4272

Kleinberg J, Lakkaraju H, Leskovec J, et al (2018) HUMAN DECISIONS AND MACHINE PREDICTIONS. Q J Econ 133:237–293. https://doi.org/10.1093/QJE/QJX032

Kleinberg J, Mullainathan S, Raghavan M (2016) Inherent Trade-Offs in the Fair Determination of Risk Scores. ArXiv ID: 1609.05807v2

Kostka G (2019) China’s social credit systems and public opinion: Explaining high levels of approval: https://doi.org/101177/1461444819826402 21:1565–1593. https://doi.org/10.1177/1461444819826402

Kroll J, Huey J, Barocas S, et al (2017a) Accountable Algorithms. Univ PA Law Rev 165:

Kroll JA (2018) The fallacy of inscrutability. Philos Trans R Soc A Math Phys Eng Sci 376:. https://doi.org/10.1098/RSTA.2018.0084

Kusner MJ, Loftus JR (2020) The long road to fairer algorithms. Nature 578:34–36. https://doi.org/10.1038/D41586-020-00274-3

Kwet M (2019) Digital Colonialism is Trheatening the Global South. Al Jazeera

Lamo de Espinosa E (2004) Bajo puertas de fuego: el nuevo desorden internacional. Taurus

Lang PJ (2004) Fear reduction and fear behavior: Problems in treating a construct. In: Research in psychotherapy. American Psychological Association, pp 90–102

Lee MSA, Floridi L (2020) Algorithmic Fairness in Mortgage Lending: From Absolute Conditions to Relational Trade-Offs. SSRN Electron J. https://doi.org/10.2139/SSRN.3559407

Lei YW (2021) Delivering Solidarity: Platform Architecture and Collective Contention in China's Platform Economy: https://doi.org/101177/0003122420979980 86:279–309. https://doi.org/10.1177/0003122420979980

Luke.S. (2013) Essentials of Metaheuristics. Lulu, second edition, 2013. Available for free at http://cs.gmu.edu/↤sean/book/metaheuristics/.

Lyotard J-F (1996) Moralidades Posmodernas. Tecnos, Madrid

Maass A (1999) Linguistic Intergroup Bias: Stereotype Perpetuation Through Language. Adv Exp Soc Psychol 31:79–121. https://doi.org/10.1016/S0065-2601(08)60272-5

Maclure J (2021) Correction to: AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind. Minds Mach 2021 314 31:637–637. https://doi.org/10.1007/S11023-021-09576-5

MacMahon C (2016) Reasonableness and fairness. A historical theory. Cambridge University Press

Macpherson CB (2005) La Teoria politica del invididualismo posesivo. De Hobbes a Locke. Editorial Trotta

Maffie MD (2020) The Perils of Laundering Control through Customers: A Study of Control and Resistance in the Ride-hail Industry: https://doi.org/101177/0019793920972679. https://doi.org/10.1177/0019793920972679

Mántaras RL (2017) Ética en la inteligencia artificial. Investigación y Ciencia 491: 49

Mántaras RL (2020) El traje nuevo de la Inteligencia Artificial. Investigación y Ciencia 526: 50 - 59

Manzini T, Chong LY, Black AW, Tsvetkov Y (2019) Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. NAACL HLT 2019 - 2019 Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol - Proc Conf 1:615–621. https://doi.org/10.18653/V1/N19-1062

Maquiavelo N (2012) el principe, Traducción. Espasa Libros

Marks IM (1969) Fears and Phobias. Elsevier

Martin K, Waldman A (2022) Are Algorithmic Decisions Legitimate? The Effect of Process and Outcomes on Perceptions of Legitimacy of AI Decisions. Journal of Bussiness Ethics. https://doi.org/10.1007/S10551-021-05032-7

Martínez-Plumed F, Contreras-Ochando L, Ferri C, et al (2019) CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Trans Knowl Data Eng 33:3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

Merler M, Ratha N, Feris RS, Smith JR (2019) Diversity in Faces

Mikolov T, Chen K, Corrado G, Dean J (2013a) Efficient Estimation of Word Representations in Vector Space

Mikolov T, Sutskever I, Chen K, et al (2013b) Distributed Representations of Words and Phrases and their Compositionality. ArXiv: 1310.4546

Miller C, Coldicott R (2019) People, power and technology, the tech workers' view. In: Doteveryone. https://doteveryone.org.uk/report/workersview/. Accessed 3 Dec 2021

Mittelstadt BD, Allo P, Taddeo M, et al The ethics of algorithms: Mapping the debate. https://doi.org/10.1177/2053951716679679

Mökander J, Floridi L (2021) Ethics-Based Auditing to Develop Trustworthy AI. Minds Mach 31:323–327. https://doi.org/10.1007/S11023-021-09557-8

Montes N. and Sierra C. (2021). Value-guided synthesis of parametric normative systems. In Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '21, page 907–915, Richland, SC, 2021. International Foundation for Autonomous Agents and Multiagent Sys- tems. (Best paper award finalist).

Montes N. and Sierra C. (2022) Synthesis and properties of optimally value-aligned normative systems. Under review.

Moore P V. (2019) Quantified self in precarity : work, technology and what counts. Routledge

Morgado I (2017) Emociones Corrosivas. Editorial Planeta

Morley J, Elhalal A, Garcia F, et al (2021a) Ethics as a Service: A Pragmatic Operationalisation of AI Ethics. Minds Mach 31:239–256. https://doi.org/10.1007/S11023-021-09563-W

Morley J, Floridi L, Kinsey L, Elhalal A (2020) From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. Sci Eng Ethics 26:2141–2168. https://doi.org/10.1007/S11948-019-00165-5/TABLES/6

Morley J, Kinsey L, Elhalal A, et al (2021b) Operationalising AI ethics: barriers, enablers and next steps. AI Soc. https://doi.org/10.1007/S00146-021-01308-8

Morley J, Kinsey L, Elhalal A, et al (2021c) Operationalising AI ethics: barriers, enablers and next steps. AI Soc. https://doi.org/10.1007/S00146-021-01308-8

Mounk Y (2017) The Age of Responsibility: luck, choice and the welfare state. Cambridge University Press

Mumford L (1997) Técnica y civilización. Alianza

Nadeem M, Bethke A, Reddy S (2020) StereoSet: Measuring stereotypical bias in pretrained

language models. 5356–5371

Narayan D, Petesch P (eds) (2002) Voices of the Poor. From Many Lands. Oxford University Press & The World Bank

Navarro EM (2002) Aporofobia. Glosario para una Soc. Intercult. 17–23

Newton D (2021) AI is infiltrating college, from admissions to teaching to grading. https://eu.usatoday.com/story/news/education/2021/04/26/ai-infiltrating-college-admissions-teaching-grading/7348128002/. Accessed 26 Feb 2022

Nikhil Garga, Londa Schiebingerb, Dan Jurafskyc JZ (2018) Word embeddings quantify 100 years of gender and ethnic stereotypes. 12:. https://doi.org/10.1073/pnas.1720347115

Nunn H, Biressi A (2009) The undeserving poor. Soundings 41:

Nussbaum MC (2012) Creating Capabilities. Harvard University Press, Cambridge, Massachusetts and London, England

O'Neal C (2016) Weapons of Math Destruction. Penguin Random House

OECD (2018) A Broken Social Elevator? How to Promote Social Mobility. OECD

Ortega Gasset J (1942) Ideas y Creencias. Ediciones de la Revista de Occidente, Madrid

Osnos E. (2014). Age of Ambition: Chasing For- tune, Truth, and Faith in the New China. Farrar, Strauss and Giroux, May 2014.

Ostry JD, Berg A, Tsangarides C (2014) Redistribuion, inequality and growth

Paolini S, White F, Tropp L, et al (2021) Intergroup contact research in the 21st century. Lessons learned and forward progress if we remain open. Journal of Social Issues

Pennington J, Socher R, Manning CD (2014) GloVe: Global Vectors for Word Representation. In: Empirical Methods in Natural Language Processing (EMNLP). pp 1532–1543

Peters D (2019) Beyond Principles: A Process for Responsible Tech | by Dorian Peters | The Ethics of Digital Experience | Medium. https://medium.com/ethics-of-digital-

experience/beyond-principles-a-process-for-responsible-tech-aefc921f7317. Accessed 3 Dec 2021

Pettigrew TF (2020) Contextual social psychology : reanalyzing prejudice, voting, and intergroup contact. 271

Pettigrew TF (2021). Advancing intergroup contact theory: Comments on the issue's articles. Journal of Social Issues, 77(1):258–273.

Piketty T (2014) EL CAPITAL EN EL SIGLO XXI. Fondo de Cultura Económica, Bogotá

Piketty T (2020) Capital and Ideology. Ediciones Deusto

Piketty T, Saez E, Gabriel Zucman, et al (2018) Distributional National Accounts: Methods and Estimates for the United States. Q J Econ 133:553–609. https://doi.org/10.1093/QJE/QJX043

Platón (2013) La República o el Estado. Austral

Poell T, Nieborg D, van Dijck J (2019) Platformisation. Internet Policy Rev 8:. https://doi.org/10.14763/2019.4.1425

Poitras L (2014) Citizenfour. https://www.filmaffinity.com/es/film740797.html

Qiu JL (2017) Goodbye iSlave : a manifesto for digital abolition. University of Illinois Press

Radford A, Wu J, Child R, et al (2019) Language Models are Unsupervised Multitask Learners. Computer Science

Raghu M, Blumer K, Corrado G, et al (2019) The Algorithmic Automation Problem: Prediction, Triage, and Human Effort. ArXiv ID: 193.122220v1

Rahman HA, Valentine MA (2021) How Managers Maintain Control Through Collaborative Repair: Evidence from Platform-Mediated "Gigs." https://doi.org/101287/orsc20211428 32:1300–1326. https://doi.org/10.1287/ORSC.2021.1428

Ratnaparkhi TS, Tandasi A, Saraswat S (2021) Face Detection and Recognition for Criminal

Identification System,. Xplore, 11th Int Conf Cloud Comput Data Sci Eng 773–777

Ravallion M (2004) Pro-Poor Growth: A Primer. Development Research Group. World Bank

Rawls J (1971) A Theory of Justice. Oxford University Press, Oxford

Reicher S (2007) Rethinking the paradigm of prejudice. South African J Psychol 37:820–834. https://doi.org/10.1177/008124630703700410

Reis E, Moore M, Clarke G, et al (2005) Elite perceptions on poverty and inequality. Zed Books, London

Ridgeway CL, Smith-Lovin L (1999) The gender system and interaction. Annu Rev Sociol 25:191–216. https://doi.org/10.1146/ANNUREV.SOC.25.1.191

Roberts H, Cowls J, Hine E, et al (2021) Achieving a 'Good AI Society': comparing the aims and progress of the EU and the US. SSRN Electron J. https://doi.org/10.2139/SSRN.3851523

Rousseau J-J (1923) Discurso sobre el origen de la desigualdad entre los hombres. Calpe

Rudinger R, Naradowsky J, Leonard B, Durme B Van (2018) Gender Bias in Coreference Resolution. NAACL HLT 2018 - 2018 Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol - Proc Conf 2:8–14. https://doi.org/10.18653/V1/N18-2002

Sampedro J (1972) Conciencia del subdesarrollo, Primera ed. Salvat Editores, Alianza Editorial, Estella, Spain

Sampson, O., & Chapman M (2021) AI Needs an Ethical Compass. This Tool Can Help. | ideo.com. https://www.ideo.com/blog/ai-needs-an-ethical-compass-this-tool-can-help. Accessed 4 Dec 2021

Sandel M (2020a) Hacia una política del bien común | Predicciones sobre el Coronavirus en EL PAÍS. In: El Pais. https://elpais.com/especiales/2020/coronavirus-covid-19/predicciones/hacia-una-politica-del-bien-comun/. Accessed 9 Jan 2022

Sandel MJ (2020b) The tyranny of merit. Penguin Random House

Sandel MJ (2013) What money can't buy. The moral limits of markets. Penguin Books

Sallila S. (2010). Using microsimulation to optimize an income transfer system towards poverty reduction. Journal of Artificial Societies and Social Simulation, 13(1), 2010.

Sap M, Gabriel S, Qin L, et al (2020) Social Bias Frames: Reasoning about Social and Power Implications of Language. 5477–5490. https://doi.org/10.18653/V1/2020.ACL-MAIN.486

Schick T, Schütze H (2020) It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. 2339–2352. https://doi.org/10.18653/v1/2021.naacl-main.185

Schopenhauer A (1993) Los dos problemas fundamentales de la ética. Siglo XXI

Schwartz R, Dodge J, Smith NA, Etzioni O (2019) Green AI. Commun ACM 63:54–63. https://doi.org/10.1145/3381831

SCMP Research (2020) China AI Report. https://www.worldscientific.com/page/china-ai-report. Accessed 5 Feb 2022

Sen A (2001) Development as freedom. Oxford University Press

Sierra C, Osman N, Noriega P, et al (2021) Value alignment: a formal approach

Smith A (2020) Adam Smith La riqueza de las naciones. Traducción Carlos Rodríguez Braun. Biblioteca Nueva

Smith A (2016) The Theory of Moral Sentiments. Enhanced Media Publishing

Smith N.P. and Deranty (2011) Labour: Work and the Social Bond. Brill.

Solaiman I, Brundage M, Clark J, et al (2019) Release Strategies and the Social Impacts of Language Models. ArXiv ID: 1908.09203

Srnicek N (2016) Platform capitalism. Wiley

Standing G (2011) The Precariat. The Precariat. https://doi.org/10.5040/9781849664554

Stark D, Pais I (2020) Algorithmic management in the platform economy. Sociologica 14:47–72. https://doi.org/10.6092/ISSN.1971-8853/12221

Talmor A, Yoran O, Le Bras R, et al (2021) CommonsenseQA 2.0: Exposing the Limits of AI through Gamification. Thirty-fifth Conf Neural Inf Process Syst Datasets Benchmarks Track (Round 1), 2021

Taylor C (1931) Multiculturalism and "the politics of recognition." Princeton University Press

Terzis P (2020) Onward for the freedom of others: Marching beyond the AI ethics. FAT* 2020 - Proc 2020 Conf Fairness, Accountability, Transpar 220–229. https://doi.org/10.1145/3351095.3373152

The World Bank. (2020) Reversals of Fortune. Techreport.

Tortosa J (2001) El juego global. Maldesarrollo y pobreza en el capitalismo mundial. Icaria, Barcelona

Townson S (2020) AI Can Make Bank Loans More Fair. In: Harv. Bus. Rev. https://hbr.org/2020/11/ai-can-make-bank-loans-more-fair. Accessed 5 Feb 2022

Tsamados A, Aggarwal N, Cowls J, et al (2021a) The ethics of algorithms: key problems and solutions. AI Soc 2021 1:1–16. https://doi.org/10.1007/S00146-021-01154-8

Tsamados A, Aggarwal N, Cowls J, et al (2021b) The ethics of algorithms: key problems and solutions. AI Soc 2021 1:1–16. https://doi.org/10.1007/S00146-021-01154-8

Ugwudike P (2021) AI audits for assessing design logics and building ethical systems: the case of predictive policing algorithms. AI Ethics 2021 1:1–10. https://doi.org/10.1007/S43681-021-00117-5

United Nations. Department of Economics and Social Affairs (2020) World Social Report

Vakkuri V, Kemell KK (2019) Implementing AI Ethics in Practice: An Empirical Evaluation of the RESOLVEDD Strategy. Lect Notes Bus Inf Process 370 LNBIP:260–275.

https://doi.org/10.1007/978-3-030-33742-1_21

Vakkuri V, Kemell KK, Jantunen M, Abrahamsson P (2020) “This is Just a Prototype”: How Ethics Are Ignored in Software Startup-Like Environments. Lect Notes Bus Inf Process 383 LNBIP:195–210. https://doi.org/10.1007/978-3-030-49392-9_13

Vallas S, Schor JB (2020) What Do Platforms Do? Understanding the Gig Economy. https://doi.org/101146/annurev-soc-121919-054857 46:273–294. https://doi.org/10.1146/ANNUREV-SOC-121919-054857

Vallès-Peris N, Domènech M (2021) Caring in the in-between: a proposal to introduce responsible AI and robotics to healthcare. AI Soc. https://doi.org/10.1007/S00146-021-01330-W

van Dijk J (2020) The Network Society. Sage Publications, London

van Nood R, Yeomans C (2021) Fairness as Equal Concession: Critical Remarks on Fair AI. Sci Eng Ethics 27:. https://doi.org/10.1007/S11948-021-00348-Z

Vinuesa R, Azizpour H, Leite I, et al (2020) The role of artificial intelligence in achieving the Sustainable Development Goals. Nat Commun 2020 111 11:1–10. https://doi.org/10.1038/s41467-019-14108-y

von Eschenbach WJ (2021) Transparency and the Black Box Problem: Why We Do Not Trust AI. Philos Technol 34:1607–1622. https://doi.org/10.1007/S13347-021-00477-0

Vynck G De (2021) Autonomous weapons already exist and are playing a role on battlefields like Libya and Armenia - The Washington Post. In: Washington Post. https://www.washingtonpost.com/technology/2021/07/07/ai-weapons-us-military/. Accessed 5 Feb 2022

Watson DS, Krutzinna J, Bruce IN, et al (2019) Clinical applications of machine learning algorithms: Beyond the black box. BMJ 364:. https://doi.org/10.1136/BMJ.L886

West SM, Whittaker M, Crawford K (2019) DISCRIMINATING SYSTEMS Gender, Race, and Power in AI. AI Now Institute.

White House (2016) Big Risks, Big Opportunities: the Intersection of Big Data and Civil Rights

Whittlestone J, Nyrup R, Alexandrova A, Cave S (2019) The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. AIES 2019 - Proceedings of the 2019 AAAI / ACM Conference on AI, Ethics and Society

Wilson WJ (1987) The truly disadvantaged: The inner city, the underclass, and public policy. Univ Chicago Press

Wilson WJ (1996) When Work disappears. The World of the New Urban Poor. Random House Inc

Wong PH (2020) Cultural Differences as Excuses? Human Rights and Cultural Values in Global Ethics and Governance of AI. Philos Technol 2020 334 33:705–715. https://doi.org/10.1007/S13347-020-00413-8

World Economic Forum (2022) Advancing Digital Agency: The Power of Data Intermediaries

Yapa L (2002) How the discipline of geography exacerbates poverty in the Third World. Futures 34:33–46

Young M (1964) The Rise of the Meritocracy (Classics in Organization and Management Series). Routledge

Zajko M (2021) Conservative AI and social inequality: conceptualizing alternatives to bias through social theory. AI Soc 36:1047–1056. https://doi.org/10.1007/S00146-021-01153-9

Zetterholm MV, Lin Y, Jokela P (2021) Digital Contact Tracing Applications during COVID-19: A Scoping Review about Public Acceptance. Informatics 2021, Vol 8, Page 48 8:48. https://doi.org/10.3390/INFORMATICS8030048

Zhao J, Khashabi D, Khot T, et al (2021) Ethical-Advice Taker: Do Language Models

Understand Natural Language Interventions? 4158–4164. https://doi.org/10.18653/v1/2021.findings-acl.364

Zhao J, Wang T, Yatskar M, et al (2018) Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. ACL Anthology. 15–20

Zuboff S (2019) The age of surveillance capitalism. Profile Books