CY. center for
algorithmic
transparency

Document Title	**Algorithm Watchdog**
Project Title and acronym	Cyprus Center for Algorithmic Transparency (CyCAT)
H2020-WIDESPREAD-05-2017-Twinning	Grant Agreement number: 810105 — CyCAT
Deliverable No.	D6.6
Work package No.	WP6
Work package title	Inter-institutional Networking
Authors (Name and Partner Institution)	Lena Podoletz (UEDIN) Michael Rovatsos (UEDIN)
Contributors **(Name and Partner Institution)**	
Reviewers	Monica L. Paramita (USFD)
Status **(D: draft; RD: revised draft; F: final)**	F
File Name	D6.6_Algorithm_Watchdog
Date	16 December 2021

Draft Versions - History of Document				
Version	**Date**	**Authors / contributors**	**e-mail address**	**Notes / changes**
V1.0	01/09/2021	Lena Podoletz Michael Rovatsos	lena.podoletz@ed.ac.uk michael.rovatsos@ed.ac.uk	Initial draft
V2.0	27/10/21	Lena Podoletz	lena.podoletz@ed.ac.uk	Writing up the deliverable
V3.0	3/11/21	Michael Rovatsos	Michael.Rovatsos@ed.ac.uk	Reviewed and contributed to write-up
V4.0	11/11/21	Lena Podoletz	lena.podoletz@ed.ac.uk	Added outputs
V5.0	25/11/21	Michael Rovatsos	Michael.Rovatsos@ed.ac.uk	More outputs added
V6.0	30/11/21	Lena Podoletz	lena.podoletz@ed.ac.uk	Write-up
V7.0	30/11/21	Monica L. Paramita	m.paramita@sheffield.ac.uk	Reviewed the deliverable
V8.0	14/12/21	Lena Podoletz	lena.podoletz@ed.ac.uk	Made suggested edits
V9.0	15/12/21	Michael Rovatsos	Michael.Rovatsos@ed.ac.uk	Finalized

Abstract

This deliverable describes the activities that were carried out in the process of preparing the White Paper on the Algorithm Watchdog. This includes drafting the structure of an Algorithm Watchdog organisation, conducting empirical research on a selected case study and organising events to validate the research results. The final version of the White Paper describing the Algorithm Watchdog is included as an Appendix.

Keyword(s): Algorithm Watchdog, recruitment, transparency, fairness, bias, stakeholder interviews, expert workshops, policy solutions

Table of contents

1. Executive summary

This deliverable provides our vision of an algorithmic watchdog - an independent, non-profit entity that provides a transparent service for citizens' complaints investigation regarding harms caused by algorithmic systems, operates a public portal to track such disputes and their resolution, offers expert policy advice, publishes recommendations and guidelines based on their investigations and case studies and builds a library of past cases to document and advance public debate regarding the societal implications of algorithms. It also describes the activities that were carried out as part of the work package and, in appendices, contains a White Paper on Algorithm Watchdogs and the analysis of the stakeholder interviews we have conducted.

2. Introduction

In recent years we have seen a rapid increase in the use of products and services that apply algorithmic decision-making processes both in the public and the private sector. The speed of the evolution of new, never before seen technologies makes it almost impossible for generic and sectoral legal regulations to keep up with new developments. Many national and international, private and public organisations have started to recognise and acknowledge the potential problems systems that use algorithmic decision-making processes might cause to individual citizens, groups, and society as a whole. Many of these organisations have released statements, guidelines and recommendations on what they consider the most serious concerns and issues in the field of algorithmic decision-making.

The potentially global scale of the problem and the breadth of application areas where issues may arise make it very difficult for these efforts to provide concrete solutions, therefore the recommendations articulated in most of these papers remain on the level of general requirements for ethical and trustworthy technologies. Despite these difficulties, provisional solutions to the core issues are being drawn up, for example by the EU's High-Level Expert Group on Artificial Intelligence Policy and Investment Recommendations on Trustworthy AI[1], AccessNow's Human Rights in the Age of Artificial Intelligence[2], AI4People's Ethical Framework for a Good AI Society[3] or The Law Society of England and Wales 2019 Report on Algorithms in the Criminal Justice System[4]. The proposed solutions cover key issues like lawfulness, accountability, transparency, liability, sustainability and contain guidelines and more concrete recommendations regarding ethical/trustworthy AI and the regulation of algorithmic decision making. Some organisations have also compiled lists and assessments of existing and suggested regulatory solutions regarding artificial intelligence, such as AccessNow's Mapping Regulatory Proposals for Artificial Intelligence in Europe [5] and Algo:Aware's State of the Art Report[6]. It seems, however, that recommendations regarding the oversight and monitoring of algorithmic decision-making systems are missing from the landscape.

Following the spirit of these guidelines and documents we suggest that algorithmic decision-making systems require some form of oversight in order to avoid and mitigate unintended consequences such as discrimination, the use of unfair and biased systems and in order to provide a higher level of transparency, accountability and possibility to appeal against algorithmic decisions. We recommend

[1] EU High-Level Expert Group on AI (2019) Policy and Investment Recommendations on Trustworthy AI

[2] Access Now (2018) Human Rights in the Age of Artificial Intelligence

[3] AI4People (2018) An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations

[4] The Law Society (2019) Algorithms in the Criminal Justice System. The Law Society of England and Wales.

[5] Access Now (2018) Mapping Regulatory Proposals for Artificial Intelligence in Europe

[6] Algo:aware (2018) State of the Art Report: Algorithmic Decision-Making

that an **Algorithm Watchdog** be able to provide the necessary oversight and other important functions in order to mitigate against the risks of algorithmic decision-making systems and ensure they are deployed to the benefit of society and the economy. Our vision of a working Algorithm Watchdog consists of establishing an independent nonprofit body that will monitor claims from citizens and organisations regarding alleged harms of real-world, deployed products and services that rely on algorithmic components.

3. Description of activities

In order to map out the landscape, to create our recommendation on an Algorithm Watchdog and to validate our proposed structure and methodology we have conducted the following activities:

a) Desk research on Watchdog-type organisations and creating an initial proposal for a Watchdog
b) Narrowed down the problem and selecting an area of use for algorithmic decision-making systems to trial the research investigation activity for Watchdogs
c) Conducted the research investigation of the selected area: algorithmic decision-making in recruitment
    i) Desk research to map out the landscape (See Appendix 1)
    ii) Conducted stakeholder interviews (See Appendix 2)
    iii) Organised expert workshops to validate findings from stakeholder interviews, to understand the challenges further and to scope out possible solutions (See Appendix 3)
        1) Workshop 1. Presenters: Michael Rovatsos (School of Informatics, University of Edinburgh), Lena Podoletz (School of Informatics, University of Edinburgh)
        2) Workshop 2. Presenters: Monica Paramita (Information School, University of Sheffield), Frank Hopfgartner (Information School, University of Sheffield), Michael Rovatsos (School of Informatics, University of Edinburgh)
d) Refined our proposal for Algorithm Watchdogs
e) Organised an open event to present our proposal and to gather further insight
    i) Presentations: Michael Veale (University College London), Michael Rovatsos (School of Informatics, University of Edinburgh)
    ii) Discussants: Gemma Galdon Clavell (Eticas Research and Consulting, Barcelona), Shannon Vallor (Centre for Technomoral Futures, University of Edinburgh), Ansgar Koene (Ernst and Young LLP), Burkhard Schaefer (School of Law, University of Edinburgh)
f) Finalised our White Paper on Algorithm Watchdogs (See Appendix 4: Rovatsos, M., Podoletz, L., Bogina, V. (2021) Algorithm Watchdog: A CyCAT White Paper)

4. Events organised

Date and platform	Title of event	Type of event	Aim of event	Number of participants
17.03.2021, Online (Zoom)	Algorithmic Decision-Making in Recruitment and HR	Expert workshop	To validate results of stakeholder interviews and to gather insight on our proposed model of auditing algorithms	22
19.05.2021, Online (Zoom)	Algorithm Watchdog Workshop II.	Expert workshop	To gather further insight on using algorithm auditing in recruitment and to test whether the model would work in another area (search engine bias)	16
22.09.2021, Online (Zoom)	Watching Algorithms under the Emerging EU Regulatory Framework for AI	Open event	To present our proposed model of an Algorithm Watchdog, to gather further insight on potential challenges and to map out how our proposed model could fit with the emerging EU regulatory framework on AI	63

5. Results

As a result of the activities carried out in this work package we have engaged with stakeholders and experts, organised workshops and an open event accessible for the members of the public as well, proposed framework for an Algorithm Watchdog, developed an initial methodology for carrying out investigations of problematic algorithmic systems, tested the method against the use case of algorithmic systems used in job recruitment and produced a White Paper. The desk research and the stakeholder interviews we conducted provided us with a map of the challenges found in the job recruitment landscape when it comes to algorithmic systems. In the desk research we examined existing organisations and bodies that fulfil our definition of a watchdog organisation to scope typical activities, practices, methods of operation, workflows, powers and transparency-related measures. Our stakeholder interviews provided insight to the key problems related to the use of algorithmic systems in hiring. We learned that the participants seemed most concerned about the lack of transparency, biases in algorithmic systems, the limitation of currently available tools and how this may generate skewed outcome of recommended candidates, the fact that there are no incentives for vendors do disclose information related to these problems, the limitations of relevant knowledge on the side of recruiters and applicants, the lack of accountability for decisions when it comes to vendors and developers, and the lack of clear guidelines and regulations on the development and use of algorithmic decision-making systems in hiring. As a result of our expert workshops we refined our proposed structure for an algorithm watchdog. Our final public event resulted in a set of open questions which will require further study. During our stakeholder and expert engagement we have built a small community of experts that participated in our events and provided feedback and insight on our activities. Finally, we have organised a public event to share our findings and summarised our work and results in a CyCAT White Paper.

The work conducted on the design of an Algorithm Watchdog was linked to other activities undertaken as part of the CyCAT project. When conducting the desk-research and conceptualising the dissemination work in the forms of archives and reports the Algorithm Watchdog would undertake, we relied on findings from Work Package 4 related to educating the general public and end users on algorithmic transparency and bias. This aspect of CyCAT's work is also key when it comes to citizen engagement with submitting complaints about problematic systems. The research that was undertaken in Work Package 5 on identifying and visualising bias in search engines was used in one of our workshops as an example use-case, where the expert participants provided insight on how the proposed workflow of the Algorithm Watchdog could be applied to investigating search engine bias. As part of our efforts in inter-institutional network-building, we placed an emphasis on inviting the participants of the Dagstuhl Seminar, the STSEs and the CyCAT Winter School to be contributors on our expert workshops in order to strengthen connections with the informal network of experts which was formed as a result of the CyCAT projects' activities.

Appendix 1: Exploring existing examples of Watchdog-type organisations

For the purpose of this work, in general, we define a watchdog as an organisation or person that performs its tasks at least partly to investigate illegal (or at the very least unethical) practices in one or more areas. In this short introductory section, we will only cover organisational examples of watchdog activities as this is the form of a watchdog we are looking to create a blueprint of, but we recognise that it is possible for individuals, such as academics, journalists, bloggers or individual activists to perform watchdog-type activities as well. We have identified three main subsets of watchdog type organisations based on their powers and activities: 1. Official bodies, 2. Unofficial bodies (e.g. NGO-s), 3. Groups of experts. Below we will illustrate the diverse nature of each subtypes of watchdog-like organisations with a few selected examples in each case.

1. Typology

1.1. Official bodies

By official bodies we mean organisations, public bodies or other entities that were established by law or any act of central or local government and/or are performing their duties as public duties and/or hold powers by which they can mandate stakeholders within their jurisdiction to act or not to act in a certain way.

Table 1. Summary of ‘Official bodies’

Independent Office for Police Conduct[7] (England and Wales)	
General area of activities	Policing
Aims and goals	Overseeing the police complaints system
Duties, tasks and other activities	Investigating the most serious cases of police conduct (e.g. death following police action) Handling requests for review/appeals against how a complaint was handled Setting standards for handling complaints for police forces, publishing guidelines Publishing reports on cases handled Releasing statutory guidelines for police forces to help with compliance
Powers	No data
Relationship to central/local power and other stakeholders	Independent from police, government and interest groups
Transparency and accountability	Publishing annual reports and plans Publishing quality and service standards Possibility to request further information (e.g. on specific types of complaints)
Funding	No data
Information Commissioner’s Office[8] (United Kingdom)	
General area of activities	Information rights
Aims and goals	Upholding information rights in the public interest
Duties, tasks and other activities	Upholding information rights in the public interest (e.g. Data Protection Act, Freedom of Information Act) Registering organisations that pay the mandatory fee for processing personal information Publishing certain data of above-mentioned organisations Addressing enquiries, written concerns and complaints regarding information rights Engaging with data protection regulators from other jurisdictions and the international community

[7] https://www.policeconduct.gov.uk

[8] https://ico.org.uk

	Issuing reviews and reports
Powers	Issuing penalties for failing to pay the fee for processing personal information Issuing penalties for clear and serious breaches of legislation
Relationship to central/local power and other stakeholders	Independent
Transparency and accountability	List of actions taken (e.g. penalties issued) Publishing audits and overview reports
Funding	No data
Financial Conduct Authority[9] (United Kingdom)	
General area of activities	Financial services
Aims and goals	Protecting consumers Protecting financial markets Promoting competition
Duties, tasks and other activities	Monitoring which firms and individuals can enter the financial markets Supervising whether firms uphold the standards they regulate Intervening where firms do not follow their rules Investigating markets Taking steps to address features of markets that can inhibit effective competition
Powers	Imposing penalties (issuing fines) Stopping firms from trading (e.g. withdrawing firms' authorisation, suspending firms or individuals from undertaking regulated activities) Securing redress for consumers Making public announcements about decisions, issuing warnings and alerts about unauthorised firms
Relationship to central/local power and other stakeholders	Independent but accountable to the Treasury
Transparency and accountability	Accountable to the Treasury (and ultimately, the UK Parliament)
Funding	Funded by the regulated firms

[9] https://www.fca.org.uk

Independent Prison Monitoring Advisory Group[10] (Scotland)	
General area of activities	Independent Prison Monitors
Aims and goals	Ensuring the independence of Independent Prison Monitors
Duties, tasks and other activities	Reviewing the effectiveness of prison monitoring Contributing to the guidance of prison monitoring Reviewing the training of Prison Monitors
Powers	No data
Relationship to central/local power and other stakeholders	Independent
Transparency and accountability	No data
Funding	No data
Scottish Legal Complaints Commission[11] (Scotland)	
General area of activities	Complaints against lawyers in Scotland
Aims and goals	Providing a single point of contact for complaints against all lawyers in Scotland
Duties, tasks and other activities	Investigating and resolving complaints about service Referring complaints about conduct to the relevant professional body Overseeing complaint handling by professional bodies Advising on good complaint handling
Powers	No data
Relationship to central/local power and other stakeholders	Can recommend or order an apology, reduction/refund of fees, compensation for loss, compensation for inconvenience/distress, 'putting it right'
Transparency and accountability	Publishing budget Publishing rules, policies on complaint handling Possible to make a freedom of information request Publishing case studies Possible to appeal decision in front of the Court of Session
Funding	Funded by a levy paid by legal professionals who operate in Scotland

[10] https://www.prisonsinspectoratescotland.gov.uk/get-involved/monitoring-advisory-group

[11] https://www.scottishlegalcomplaints.org.uk

1.1.1. Examples for activities and workflows

a) Independent Office for Police Conduct

The IOPC handles two types of cases: a) police conduct cases that had been referred to them due to their seriousness, and b) cases where the individual affected by the matter feels that their complaint has not been handled properly by the police forces.[12] In both types of cases the IOPC usually conducts an independent investigation, publishes a short report, and in some cases makes concrete recommendations which are mainly about how such matters should have been handled. The IOPC also released a set of statutory guidelines[13] regarding the police complaints system in order to assist police forces with legal compliance.

In the first instance complaints regarding police conduct can be submitted either directly to the police force in question or centrally and mainly they are handled by the police force.[14] The exceptions from this are serious cases which for example resulted in a person's death where the investigation is handled by the IOPC. If the person is unhappy with how their complaint was handled by the police, they can make an appeal or request a review from the IOPC as well as other bodies.[15] Based on the investigation reports on the IOPC website the IOPC uses footages, documents and other records, hearings and testimonies from involved parties and witnesses to establish what has happened and decide about appropriate actions (but it is the police force who carries out the disciplinary actions).[16] [17]

b) Information Commissioner's Office

A key watchdog-type activity the ICO is undertaking is handling complaints about organisations' practices regarding information rights. These include topics such as nuisance calls and messages, troubles accessing or re-using official or public information that one has asked from a public body, accessing personal information from an organisation, concerns about how an organisation handles personal information, the use of cookies or a provider's refusal to remove links to information about a person.[18]

In case of reporting a spam email for instance, the ICO will investigate the case and try to identify the organisation, check whether the organisation follows the rules that regulate direct marketing and take actions against the organisation (such as monetary penalties), if necessary.[19] Another example is the case of a provider refusing to remove search results that can have a negative effect on the person who complains. Here the ICO will evaluate whether the search result in question breaches the principles of data protection (e.g. inaccuracy or irrelevancy) and whether the public interest of continued access to the information is greater than the potential negative effect on the individual. As a result, the ICO will notify the search provider that the search result should be removed.[20]

c) Financial Conduct Authority

Amongst other activities the FCA investigates potential cases of harm and misconduct. An investigation is started when there is a reason to suspect serious misconduct (e.g. mis-selling of

[12] https://policeconduct.gov.uk/who-we-are
[13] https://policeconduct.gov.uk/complaints-reviews-and-appeals/statutory-guidance
[14] https://policeconduct.gov.uk/complaints-reviews-and-appeals/make-complaint
[15] https://policeconduct.gov.uk/complaints-reviews-and-appeals/reviews-and-appeals
[16] https://policeconduct.gov.uk/investigations/our-investigation
[17] https://policeconduct.gov.uk/investigations/what-we-investigate-and-next-steps
[18] https://ico.org.uk/make-a-complaint/
[19] https://ico.org.uk/make-a-complaint/nuisance-calls-and-messages/spam-emails/
[20] https://ico.org.uk/make-a-complaint/search-results/

unsuitable products to consumers).[21] Usually, the appointed investigators carry out scoping discussions and investigative work which can include informing the investigated firm or individual about the process and milestones, requests for relevant documents or information, and interviews with people. The FCA can share preliminary findings with the firm or individual to which they can respond. Once the FCA made an assessment of the case and, if needed, appropriate sanction, the case can end in different ways, such as by closing, resolution or in front of a Tribunal.[22]

d) Scottish Legal Complaints Commission

The SLCC specialises in handling complaints against lawyers. The usual workflow of such cases is to check whether the complaint was premature (e.g. whether the lawyer had an opportunity to 'put things right'), check the eligibility of the complaint (e.g. it was submitted within the time limit and its topic is within the scope of the SLCC), offering mediation if appropriate, carrying out a formal investigation, proposing settlement, and, if the settlement is not accepted, making a binding decision.[23]

1.2. Unofficial bodies

By unofficial bodies here we mean all organisations that do not fall under the category of 'official body' but perform watchdog-type activities such as independent investigation of illegal actions of authorities and corporations, publishing independent, expert reports on such issues or releasing guidelines about compliance with law or ethics. As these organisations do not have 'official powers', we will leave this factor out of the table below.

Table 2. Summary of 'Unofficial bodies'

American Civil Liberties Union[24]	
General area of activities	Constitutional rights and freedoms in the USA
Aims and goals	Protecting civil liberties in the USA
Tasks and other activities	Being involved in court cases (e.g. filing suits, representing people and organisations) Being involved in Supreme Court cases Advocacy in certain areas of public policy on state and federal level Publishing informative and educational content on individual rights (e.g. protesters' rights)
Relationship to central/local power and other stakeholders	Independent from government
Transparency and accountability	Publishing annual reports Publishing IRS (Internal Revenue Service) forms and audited financial statements

[21] https://www.fca.org.uk/about/enforcement/investigation-opening-criteria
[22] https://www.fca.org.uk/publication/corporate/enforcement-information-guide.pdf
[23] https://www.scottishlegalcomplaints.org.uk/your-complaint/our-process/
[24] https://www.aclu.org

Amnesty International[25]	
General area of activities	Human rights
Aims and goals	Defending human rights, freedoms, dignity, justice and truth
Tasks and other activities	Investigating and reporting abuses of human rights Educating the public on issues such as human rights, climate change or LGBTQ Mobilising the public Lobbying at governments and private companies
Relationship to central/local power and other stakeholders	Independent from government
Transparency and accountability	Publishing annual reports Publishing IRS (Internal Revenue Service) forms and audited financial statements
Innocence Project[26]	
General area of activities	Wrongful convictions in the USA
Aims and goals	Turning around wrongful convictions and exonerating the wrongly convicted people via the use of DNA testing
Tasks and other activities	Representing clients who seek post-conviction DNA testing Examining cases to determine whether DNA testing could prove innocence Raising awareness on unvalidated forensic science disciplines and issues around eyewitness testimonies
Relationship to central/local power and other stakeholders	Independent from government
Transparency and accountability	No data

1.3. Expert Groups

This subset of watchdogs includes groups of experts (official or unofficial) without formal powers who purely focus on releasing case studies, guidelines, recommendations but do not engage in activism, legal representation or other direct involvement in concrete cases. We separated these from the other types based on the fact that their work is mainly theoretical in nature and concentrates on

[25] https://www.amnesty.org/en/who-we-are/; https://www.amnesty.org.uk
[26] https://innocenceproject.org

analysis of cases and problems but not on engaging with complaints or handling real-life cases. The reason for including these groups as watchdogs-type organisations is that through use cases they point out potential real-life problems and breaches of legislation.

Table 3. Summary of 'Groups of experts'

Algorithm Watch[27]	
General area of activities	Algorithmic decision-making processes
Aims and goals	Evaluating algorithmic decision-making processes that are used to predict/prescribe human action or to make automated decisions
Tasks and other activities	Analysing and explaining the effects of algorithmic decision-making processes Linking experts in the field Developing ideas and strategies for intelligibility
Relationship to central/local power and other stakeholders	Independent from government
Transparency and accountability	Adhering to the governance reporting framework 'Initiative for a Transparent Civil Society' designed by TI Germany (this includes publishing annual financial statements and auditor's statements)
Princeton: Dialogues on AI and Ethics[28]	
General area of activities	AI and ethics
Aims and goals	Providing interdisciplinary insight into the intersection of AI, ethics and policy
Tasks and other activities	Creating fictional case studies from existing examples and discussing specific problematic questions regarding the example
Relationship to central/local power and other stakeholders	No data
Transparency and accountability	No data

1.4. Conclusions

The tables and example workflows described above illustrate the diverse nature of organisations that carry out watchdog activities as well as the diversity of the activities themselves. During our research it became apparent that watchdog organisations tend to have a specific focus for their activities, which usually consists of a single domain (some domains are much larger than others) and thus the activities of the organisation are limited to the domain. It also seems important – at least in the examples we have looked at – for these organisations to declare independent status and operate outside of governmental, political and industrial structures. Relating to this, transparency and accountability is

[27] https://algorithmwatch.org/en/

[28] https://aiethics.princeton.edu

also a crucial question. At least some levels of transparency seemed to be present with all examples we have looked at but when it came to accountability, information on this was mainly present in the case of official bodies only. This, of course to some extent is not unexpected as official bodies mostly perform public duties and/or are using public resources.

Appendix 2: Stakeholder interviews

The broad goal of the study was to gain insight on the use of algorithmic decision-making systems in the field of recruitment and to gather knowledge on how this might influence the presence of biases and discrimination in hiring. The aim was to explore the awareness of such issues on the part of organisations and those who work with such algorithms, to discover the challenges they may face during their work and to gain insight on their knowledge of and experiences with the potential benefits and downsides of the applications of such tools when it comes to bias and discrimination.

1. Methodology

We conducted six semi-structured, exploratory interviews with different stakeholders in order to gain insight on how algorithmic decision-making impacts the everyday practices and operations of these groups.[29] We selected our participants using purposive interviewing where we were aiming for a set of respondents who were drawn from some of the key stakeholders in the field of recruitment. Ethics clearance was applied for and was granted by the School of Informatics at the University of Edinburgh.[30] The data collection took place between October-December 2020. The interviews were conducted mainly using a set of predetermined questions tailored to the specific profile of the stakeholder, however, the semi-structured nature of the process allowed the exploration of any relevant topics that emerged during the discussion. For interviewing we used video conferencing platforms, such as Microsoft Teams and each interview lasted for approximately one hour, except for one which lasted for thirty minutes (due to time constraints on the side of the participant). With the explicit consent of the participants the audio of the interviews was recorded and partially transcribed. The quotes from the interviews have been edited only where it was necessary for clarity or anonymity. As interviews are generally not a research method which are suitable for collecting representative data and in this case it did not seem feasible to have a representative sample in size, the insights gathered from the conversations were used to identify and gain a deeper understanding of some key issues major stakeholders face in the changing landscape of recruitment regarding the recent introduction and increasing popularity of algorithmic decision-making systems. We have conducted thematic analysis of the interviews where we have identified key themes. Below we will describe each of these, illustrated with direct quotes and initial analysis.

During the analysis of the interviews it became apparent that amongst all the different themes that emerged from the discussions, there was an overarching theme: trust. The issue of trust seemed to be connected to all the other themes and thus we have decided to introduce it as an overarching theme that has an implication for all the other themes.

2. Analysis of interviews

2.1. Overarching theme: Trust in algorithmic decision-making tools

[29] We conducted interviews with the following groups of stakeholders: recruiter, large-scale employer, advocacy group.

[30] The study was certified according to the Informatics Research Ethics Process by the University of Edinburgh School of Informatics (RT #5205, Application Reference Number 27409).

One of the most important insights emerging from the interviews was the crucial role trust plays in the recruitment process. It seems that trust/distrust in these tools can affect the whole of hiring. From the interviews we have identified three major types of trust: 1. The trust the employer puts in the recruiter, 2) The trust the user of an algorithmic tool (e.g. a recruiter or an employer) puts in the tool, developer and the vendor, 3) The trust the candidate (i.e. subject) puts in the recruitment process. A respondent mentioned that whether people trust a particular tool or not can be "*very context dependent*". This converges with the fact that even though our study is looking at the field of recruitment and the hiring funnel as a whole, we cannot forget that this area uses a great variety of different tools and applications in all stages of the process.

One participant stressed that sometimes candidates feel such a high level of distrust towards certain algorithmic decision-making technologies that they may not even enter the hiring process as a result:

> *"A candidate said recently on Twitter that they were looking to apply for two jobs and both of those were using HireVue video viewing. HireVue is known for AI within its video to assess character and personality traits. The person was black (...) and [they] knew that meant the algorithms may not work for [them] (...)."* (Recruiter)

As opposed to this, another respondent brought up an example of a young female employee who was working in the STEM sector. The employee mentioned to the respondent the fact that they may feel more comfortable with a machine making decisions regarding their career rather than humans who, in their area, are "*old white males*". These two examples suggest us that the opinions of users and data subjects regarding algorithmic decision-making may have a strong correlation with the context the decision is made in, what are the alternatives to the algorithmic decision, what their perception is of the particular decision-making tool/process and what type of decision-making the person believes would guarantee the most favourable outcome for them.

2.2. Theme I. Difference between human and algorithmic decision-making

The above examples introduce us to the issue of differences between human and algorithmic decision-making and the potential need to distinguish between the two when it comes to legislation and oversight. In the context of main differences between human and algorithmic decision-making in recruitment the respondents mentioned four main topics: 1) The scale of application, 2) Feedback, 3) Potential to eliminate bias, 4) Human opinions vs. algorithmic decision.

2.2.1. The scale of application

One participant stressed the fact that human decisions can be – and most likely are – biased too. However, if an HR employee is being biased and makes biased decisions this affects 'only' the number of cases a human can deal with, whereas if bias and discrimination find their way into a code, that would apply to all people who are subjected to that system:

> *"If you code something it is applied again and again. (...) [It is the] characteristic of scale."*(Member of an advocacy group)

2.2.2. Feedback

Another participant, a recruiter, emphasised that recruitment is a human process and algorithmic decision-making often misses out on this aspect of hiring. According to them this may lead to a lack of trust in a partially or fully automated recruitment process. They also mentioned the importance of real-time feedback in particular recruitment scenarios, such as interviews:

> *"With the human engagement you get some feedback from that person, from face and body language and you can respond to that. If you see that somebody is confused and does not quite get it, you can work with that. Whereas with an algorithmic decision-making system] you just have to put the information in and say what you have to say and leave it to chance from your perspective whether you get through."* (Recruiter)

A different respondent from the recruitment industry also gave us insight on the shortcomings of current algorithmic systems used in the field of recruitment:

> "*Recruiters value human contact because what can be translated onto paper into a CV is often not the full story. (...) Some candidates are very poor at writing their CVs. What we are trying to make out with the conversations is what they can bring to the table."* (Recruiter)

These perspectives suggest that, at the moment, algorithmic decision-making systems applied in recruitment may not be able to perform as well as human recruiters because certain requirements of a particular job are not easy to describe in a job description and are even harder to meet on paper by candidates in a restrictive environment. This may result in a disadvantage for candidates who do well in one hiring environment but poorly in another (such as somebody doing well with essay questions but not with a multiple-choice test or doing well with a practical exercise but not a verbal interview). One respondent suggested that one advantageous use of machine learning technology in the field of recruitment could be to tailor hiring processes to the needs of individual candidates, so each applicant has the chance to perform best in a process-type that is comfortable for them while still being able to measure suitability for the job in question. It was also mentioned by multiple participants that if each candidate has to perform in one particular type of process, that may lead to a biased result against the applicants who were not comfortable with that type of application process. Thus, this does not seem to fit with the principle of equity.

2.2.3. Bias

When it came to the question of 'tech for good' type applications (where the most important requirement for the use of technology is its positive effects on not only effectiveness or productivity but also on societal phenomena such as equity or discrimination) of algorithmic decision-making in recruitment, some participants seemed more hopeful than others. One respondent argued that even though these systems may have great potential, it appears that as of now the technology is flawed:

> *"Companies, policy makers, entrepreneurs argue that we have this great potential. Yes, we do but where does it materialise? What we see in reality is crappy systems that*

> *are based on shady science and are being used to make real decisions."* (Member of an advocacy group)

The same participant referred to a story (albeit a single-source one) of Amazon's own system for in-house promotion which ended up being biased against female employees.[31] Their concern was that if one of the top technological companies may have been 'only' able to develop such an unsuccessful system then what kind of trust can users have in other systems when it comes to bias. Another respondent emphasised the importance of knowing who makes the decision:

> *"Even though [the candidates] would know that a human has biases, I think there is an element of being more in control of the process (...). You don't know who that development team was, whether they were all white, able-bodied males or whether it was a diverse team."* (Recruiter)

This points out that there may be a connection between trust in a decision and a decision-making process and knowing – at least to an extent – who made the decision. Even though there may not be a correlation between the knowing the identity of the decision-maker and the unbiasedness of their decision, it is very possible that the perception of candidates is that if they can pinpoint who made the decision, they have better chances knowing about possible bias and discrimination against them. Somewhat differently, a third participant took the view of algorithmic bias being possible to eliminate as opposed to human bias which most likely is not:

> *"There is a dual notion that [algorithmic decision making] has the potential to this kind of bias and discrimination [and] has the potential to spiral out of control because there is a sense that it is not under human control. On the flipside, there is a sense that with the right action and the right intervention (...) it can be shaped and controlled to a greater degree than human bias because it is about the design and these things can be explicitly designed. There is a dual sense of being scary but also quite fixable."* (Member of an advocacy group)

This indicates that even though users are aware of human bias as well, they may experience algorithmic bias as more daunting as the process seems more distant in the latter case. As one of the respondents put it, there is a "*feeling of helplessness*" when it comes to algorithmic decisions. This may be caused by different factors, such as not understanding how the system works, believing that human decision-makers are more likely to consider and positively respond to additional circumstances or unforeseeable factors, believing that human decision makers are more likely to make a fair and "humane" decision or even believing that human bias may favour them in their specific case. At the same time, the theoretical possibility to be able to debias algorithmic decisions also seems to be known, even if there appears to be a certain level of doubt towards its viability.

2.2.4. Human opinions vs. algorithmic decisions

One respondent pointed out that it is an interesting question whether people place more value on human opinions or the reliability of algorithmic processes:

[31] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

> *"I wonder (...) how much that would influence, if people said 'Oh, yes, I know that is what the clever system or tool said but when I spoke to my pal, Bob, he said Linda was a great person. (...) Where people place value, I think, is a tricky thing."* (HR representative of a large-scale employer)

This is connected to the topic of human and algorithmic bias, but also raises the question of what happens when the algorithmic decision is contradictory to a human experience, opinion or decision. It seems that the explainability of both types of decisions may be able to resolve the problem as comparing the reasoning and explanation behind human and algorithmic decisions should provide further insight into the background and reasoning of both.

Another participant pointed out that it may be misleading to talk about algorithmic decisions as opposed to human decisions as they share the same origin:

> *"Let's not talk about the human in the loop, talk about the machine in the loop. (...) The machine only executes the decision that has been made by humans long before. (...) This applies to machine learning systems where people argue that they cannot really know what the machine comes up with. This is a false depiction of reality because these systems have been trained by humans as well, even when self-learning happens. There are many presuppositions that are worked into the system before it starts to detect patterns on its own. (...) It is a dangerous narrative to say that the decision was made by a machine."* (Member of an advocacy group)

2.3. Theme II. Transparency

Another important and often talked about aspect of algorithmic systems is their opacity. For fairness, we have to note that human decisions can be equally opaque, when it comes to the personal perspectives, experiences, value and belief systems, and other factors that feed into human thought processes. When it comes to the opacity of algorithmic systems, however, a respondent argued that transparency does not necessarily have to mean the full disclosure of the whole code:

> *"[We are] not arguing that all the code or databases need to be disclosed but that they need to answer a set of questions that help people who are using these systems and people who are subject to these systems understand what the logic behind these systems is: how it was tested, whether it was tested, whether it was appropriately tested, how it works, etc."* (Member of an advocacy group)

Thus, when we are discussing transparency, we may consider different levels of transparency and it is possible that a lower level than full transparency is enough for users to make an informed decision regarding the use of the system and for subjects to understand the key aspects of the decision made about them.

Transparency has two important elements when it comes to an algorithmic decision-making system. 1) Transparency about how the system was developed, tested, etc. (as seen above); 2) Transparency about the usage of algorithmic processes in the decision-making system.

It seems that the former is particularly concerning in the field of recruitment. As one participant put it:

> *"A lot of companies would buy technology on the basis of trusting an organisation that if they say that it removes bias than it will when actually it may only do things such as remove the candidate's name. (...) People think they are getting more from the software than they are actually getting at the moment." (Recruiter)*

The same participant also pointed out that this type of misinformation may be caused by the fact that employers or even recruiters may not have the necessary skills or knowledge to find out what the algorithmic tools are actually capable of:

> *"My experience [while] working with large organisations is that HR was not really the area that got the funding or the project managers were savvy enough to grill suppliers. (...) So I think it is detrimental now that people buying in to technology in recruitment because I don't think they are strict with suppliers (...). (...) There could be more policy around transparency on what software actually does and its limitations specifically. (...) It feels like people are able to sell it as something it is not. There should be an element of proving."* (Recruiter)

This suggests that transparency has a dual challenge to tackle. Firstly, there does not seem to be a satisfactory amount of information available regarding the systems used which may be due to the fact that the vendors and developers have neither a duty nor a vested interest in disclosing that type of information. Secondly, even if the necessary information was released, it seems that employers, recruiters or simply users and subjects in general would not have the resources and skills to determine whether the system is biased, fair or whether it delivers what it promises to deliver. This highlights the importance of standards and expert evaluations.

2.4. Theme III. Responsibility, accountability and appealability

Responsibility and accountability may be one of the things that sets apart algorithmic decision-making and human decision-making. Even though in some cases, for instance, when there is a human in the loop, it seems easier to pinpoint the actor that could be held accountable for a particular decision, in general, the question of responsibility and accountability appears to be more complex in cases of algorithmic decision-making. Connected to this are subjects' worries regarding accountability for decisions that concern them and their lives. As one participant put it:

> *"People may fear that they and the decisions about them would just be embedded in code and no one would want to take responsibility for them."* (Member of an advocacy group)

It seems that in recruitment the accountability structures need to be very clear and transparent when it comes to hiring decisions. What makes this issue more complex is the many ways algorithmic decisions can impact the hiring process from phrasing the job advertisement, through placing the advert to assessment and ranking of candidates. If we consider all the stages of the hiring funnel and all the different types of algorithmic decision-making systems that are involved, it becomes apparent that the question of responsibility is more complex than saying 'the employer should be held accountable for every hiring decision they make'. In the end, of course, it seems logical to place the responsibility on the employer, as they are the one making the decision. However, we also need to consider the accountability of other actors and what measures employers need in place in order to be held accountable fairly. Hiring is a process which consists of a series of decisions rather than one single decision. Even though it may look like that the last decision is the most important one (i.e. which candidate gets the job offer at the end), all the decisions that have been made previously severely affect that final decision. For instance, how is the job advert phrased, who gets to see the advert on an advertising platform, who passes the stage of initial screening, who ends up on top of the ranking, who gets interviewed and what method is chosen for the whole process. One participant illustrated this matter with an example:

> *"I've spoken to the head of [recruitment company] about technology because they are planning to bring in video technology that assesses character. (...) Asked them what would they do in terms of the 20% of society that would not be able to get through [and] that's not counting the ones who would not even try. Their view is that you can't accommodate everybody."* (Recruiter)

A matter loosely connected to responsibility is the appealability of the decision. Appealability in our context has two elements: 1) Theoretical appealability meaning there are suitable mechanisms in place for such cases (including a possible legal obligation for review on the part of the decision-maker), it is possible to change the decision after it has been made, or at the very least, seek compensation for an unlawful/unethical/unfair decision. 2) Practical appealability meaning it is possible to gather enough information about the decision for an appeal. This duality is partially resolved by the GDPR in the EU which gives an automatic right to a review by a human, in case of fully automated decision-making. However, this only applies to fully automated decision-making meaning there is no human involved in the decision-making process (inputting data does not qualify as involvement in the decision-making here). In recruitment, however frequent biased decisions may be, it seems that it may not be very common for candidates to actually go through with appeal procedures:

> *"I've never known anybody who contested a recruitment decision. (...) They will see the bias in front of them, they will know it's happened. (...) But people tend not to do anything about it because they know, ultimately they will not get the job and they tend to put their energy into finding a job that they will get. (...) It's not like when you don't get a loan but then you complain and get the loan. (...) Candidates worry that their reputation starts to get tarnished if they report a client, and other clients may not want to interview them."* (Recruiter)

This could also mean, although this is very speculative at this point, that biased or discriminatory decisions are significantly more common than the number of known cases suggests. For this reason, when it comes to algorithmic decisions, it seems important to make sure that the requirements of both theoretical and practical appealability are met, even if candidates will not necessarily opt to actually use these mechanisms.

Explainability seems to be a key factor in both responsibility and in appealability. In the recruitment context, explainability has two sides. 1) One is that the employer should be able to comprehend why certain candidates were favoured over others or how the input will influence the output. 2) The other is that the employer, as the eventual decision-maker, needs to be able to explain the decision to the candidate.

> *"You have to be able to say [to] candidates, here is the reason why you weren't successful. (...) If you have tools which say, well, actually, what the test shows was you don't handle ambiguity well (...). It is really important because you are going to be working in an area where there is going to be lots of changes and there is going to be situations where you are going to have to make decisions on not-clear evidence for example. That is what you need to be able to show the candidate."* (HR representative at a large-scale employer)

This suggests that the potential inability of the user (in our case the employer or the recruiter) to understand how the system works and how the outputs were generated could lead to further problems regarding appealability and accountability. It also decreases the chances of the employer or the recruiter identifying a biased system or decision. Consequently, this can have an impact on trust in the system.

2.5. Theme IV. Regulation, policy and oversight

It seems that there is a need for either regulation or clarification of existing regulation on this field. For example, anti-discrimination regulation should be applicable to any situation regardless whether the decision-maker is human or algorithmic. However, in practice it does not seem clear yet who bears the responsibility for fully or semi-automated decisions in this area. What should be applicable to these types of algorithmic decision-making is the prohibition of indirect discrimination. However, this type of prohibition seems to be setting "open-ended standards" rather than a clear, strict rule (Zuiderveen Borgesius 2018:19). One key regulatory challenge, however, is caused by the complexity and opacity of algorithmic decision-making systems:

> *"(...) It is quite difficult problem in terms of regulation and policy. My understanding is that both of those things (...) lag behind the technology and the use of technology itself. (...) Most policy makers who aren't technical experts (...) A lot of policy makers aren't entirely sure how these processes operate."* (Member of an advocacy group)

This suggests that experts, especially ones with strong STEM background who also have an understanding of the legal and societal aspects of algorithmic decision-making will play a key role in shaping the future of the use of such tools in recruitment and in other areas.

A participant who works as an HR representative at a large-scale employer shared how they would go about selecting a recruitment company or algorithmic tool when hiring:

> *"If I was then to select an organisation I would want to know: (...) What is your success criteria, how does that look? But equally, are you doing that in a legitimate way? Are you working to certain standards?"* (HR representative at a large-scale employer)

This suggests that if there were clear and openly available standards, it would be possible for employers to make more informed decisions about the use of algorithmic tools. It would also make it easier for candidates to contest a decision.

> *"If there was greater responsibility, then due diligence automatically follows because there is a financial risk that follows."* (Recruiter)

This quote reminds us the importance of responsibility and responsibilisation and leads us back to the arguments regarding the need for clear accountability structures. Taking responsibility can happen organically, if companies internalise the importance of diversity and unbiased decisions and how these can add value to their work and their products. However, it also makes sense to incentivise employers to act responsibly and take accountability for their hiring decisions as these have a major effect on candidates lives, potentially their family members' lives and on a larger scale, the whole of society.

3. Key insights from interviews on algorithmic decision-making in recruitment

The following key insights emerged from the analysis of the participants responses:

1. The participants were aware that both human and algorithmic decisions could be biased. Whereas algorithmic decisions – theoretically – may be easier to debias, they also come with two important differences from human decisions'. One is the scale of the applicability as algorithmic decisions can be applied on a significantly larger scale and also in different fields. The other is that the user/subject does not know the identity of the developing team who, indirectly, partake in the decision-making process as well thus making an important actor in the decision-making unknown and almost untraceable.
2. The application of algorithmic decision-making in certain hiring scenarios may not only depend on whether the system promises but also on whether recruiters and employers place more importance on a full evaluation of a candidate or on the opinion of a human referee or recruiter. It seems the individual value systems of users can play a part in whether a system is applied.
3. It seems that current algorithmic tools in recruitment are not able to grasp the amount of information that a human recruiter can when it comes to the suitability of a candidate. This of course, can change with the further development of available tools.
4. When it comes to transparency, the recruitment industry and subject face two obstacles. One is that vendors and developers do not have the duty to disclose how their systems work and they do not have any incentives to do so either (e.g. financial interest). The other is that users and subjects, most of the time, do not possess the necessary skills or resources to evaluate the aforementioned information. Thus, it seems that intervention should focus on these points.
5. The different algorithmic tools that can potentially be involved in a hiring decision makes the issue of responsibility and accountability very complex. The recruitment process is a series of decisions rather than one single decision which ends with the final hiring decision (i.e. which person(s) get the job). This requires clearly defined accountability structures throughout the hiring funnel.

6. Even though candidates may not be very likely to appeal a hiring decision, even if they thought it was biased, there still needs to be a system that meets the requirements of both theoretical and practical appealability.
7. The requirement or standards of explainability could resolve some of the issues connected to accountability and appealability. Here we need to consider two types of explainability. One is where the employer/recruiter understands the operating principles of the system (e.g. input-output relationship), and the other is that the decision later can be explained and reasoned to the respective candidate.
8. It seems that regulatory approaches, at the very least, in the form of guidelines, standards or clear interpretations of current regulations may be needed in the field of recruitment when it comes to the algorithmic decision-making tools.

4. Summary of key challenges regarding algorithmic decision-making systems in hiring

1. Lack (or very low levels) of transparency when it comes to:
    a. The development process (e.g. was the development team diverse, what fairness and other metrics were applied, where training data came from, how testing was conducted)
    b. The internal mechanics of the tool (e.g. what data it uses, how it weighs different variables)
    c. Potential system biases
2. Biases in algorithmic systems
    a. How does this compare to biases in human decisions
    b. How to detect algorithmic bias in recruitment decisions
3. Limitations of currently available tools (e.g., the set of variables they consider are very limited compared to what a human recruiter checks)
4. No incentives for vendors to disclose information related to the above concerns
5. Limitations of relevant knowledge and skills of users and subjects to evaluate any information related to the above concerns, even if it is disclosed
6. Lack of accountability for biased decisions when it comes to developers and vendors
7. Lack of clear guidelines, standards or clear interpretations of current regulations of algorithmic decision-making tools used in the field of recruitment

Appendix 3: Expert workshops

After conducting our interviews with the stakeholders, analysing the content and describing the main themes and key insights that emerged, we presented our findings in two expert workshops. In the workshops we discussed the formation of an Algorithm Watchdog and our discussions focused on two topics regarding algorithms used in recruitment: the role of regulation and technical investigation methods. Below we will describe the insights that emerged from the workshop.

1. Access to systems

One major obstructing factor in the assessment of any algorithmic systems used in recruitment is that it may be hard for a private assessor or third-party observer to gain access to the inside of the system (as opposed to a relevant authority that can demand access).

2. Targeting specific systems and/or specific tasks

If the assessor is an official authority, the scope of its duties definitely need to be defined and this will create a group of systems that are inside and a group of systems that are outside its scope. In the context of assessing a type of system or setting out recommendations for a type of system, defining the scope is necessary for a private expert group as well. One possible way to do this is to decide to focus on a system that performs a certain task at a certain stage of the recruitment process (e.g. interview assessment). It also should be determined which algorithmic systems have actual influence on the eventual hiring decision. In this context, any decision-supporting system has the possibility to influence the actual recruitment decision. For instance, even a filtering system that, in theory, can find any candidate who fits the set criteria has the potential to influence the decision as it displays the candidates in a particular order. Another way of narrowing the scope of regulation or assessment is to define the purpose of the use so even in case of systems that can be used in different contexts, it is clear which uses are covered. For example, in the case of social media advertising where the platform and the algorithmic decision-making system can be used to advertise jobs but also anything else, there could be a special set of requirements for job advertising with specific fairness criteria or assessment methods.

3. Explainability

In the context of explainability it was suggested that it is not enough to explain how a particular technological tool works, but that it also needs to be interpreted in the societal context it is applied in. In the field of recruitment this could mean that it may not be sufficient to provide a description of how a tool works and what fairness definition it uses since this needs to be interpreted in the specific context of each use case that is determined by the societal reality of the particular job advertisement process in which the tool is used. For instance, some sectors are traditionally different from another, just like construction work tends to have a significant imbalance towards male workers and candidates. These differences should be taken into account when assessing a recruitment tool on the grounds of fairness and system bias (as they may serve as a non-discriminatory explanation for imbalances in the pool of recommended/hired candidates and the existing workforce). At the same time, they should not affect the chances of any individual candidate when it comes to applying for a particular job.

In algorithmic decision-making a huge range of variables can contribute to the final decision and it is difficult to reason the combined, weighed variable. An important question to raise is whether there is a reason to think that a particular criterion – which should not be used in the assessment – can somehow be derived from certain data. For example, can the ethnicity of the applicants be pulled from the data available and is it being used to determine the suitability of a candidate? The explanations of system decisions should be detailed enough to make it clear why certain criteria were applied if the application of such criteria was the result of algorithmic decision-making as opposed to being a simple input from a human-in-the-loop. This can also cause issues when it comes to the use or purchase of tools as the lack of explainability criteria makes it possible for companies to market products with advertised properties, many of which are arbitrarily created for greater marketability. On the other hand, accuracy should not be sacrificed in return for full explainability.

A potential role for a regulator or a watchdog organisation could be to check whether the explanations provided are true as now one of the main questions is whether we trust companies or not. One way to achieve this type of transparency is to make the minutes from the meetings of hiring committees and the evaluations of candidates available for inspection and provide access to these documents to trade unions and other relevant organisations (naturally, after proper anonymisation and privacy assurance-related procedures). This level of transparency could be achieved with proper, detailed and verifiable explanations of decisions made by algorithmic systems as well.

The example of transparency of hiring decisions could also serve as a potential route for an appeal. Another way to make an appeal is to make the appeal system based around the individual skills and CV of a particular candidate where the specific case would get in front of an external reviewer. This workflow could be transferred to algorithm-based hiring decisions as well.

One external example is the financial sector in Finland where decision-makers must have evidence that they adjusted for bias. A similar model could be introduced in the case of algorithmic systems performing certain tasks in recruitment where the developer, the vendor, or simply the company that uses the system could be required to show proof of bias-adjustments and explanations for the applied fairness model. Even if this responsibility lies with the recruiter or the employer, this would incentivise them to seek out explanations and assurances from developers and vendors. This would have the potential to introduce due diligence in the application of algorithmic decision-making systems in recruitment. This evidence then could potentially be assessed by external assessors who could be either certified private assessors or authorities. One hindering factor here could be the lack of technical expertise on the side of HR representatives.

Next to the individual-level explanations that are on some level included in GDPR, there is a need for a system-level understanding of the decisions and their consequences as well. Explanations and the assessment of explanations need to include robustness as well. For instance, a robustness study of HireVue's candidate assessment technology showed that factors such as a candidate wearing/not wearing glasses, different background on the video footage or light levels had a dramatic effect on the scores of candidates. Evidence should be shown that these tools are robust against factors that should not have any relevance when making a hiring decision.

4. Testing of the systems

From the perspective of developers, one of the most important questions is whether systems should be assessed before deployment or after as both approaches have downsides. In the case of assessment after deployment, there is a real risk of exposing candidates to discriminatory results but at the same time, testing becomes easier. If one decides to assess systems before real-life use, that may protect the job applicants from discrimination and/or system bias but it is difficult to properly test a system without having enough real-life data and thus this method may give a false assessment (which, in turn, can also expose candidates to discriminatory system behaviour).

5. Bias and discrimination in online advertising

It is important to note that online advertising by itself implies a certain level of exclusion in the sense that not everybody has or chooses to have access to online content. At the same time, if this is compared to more traditional forms of job advertisement, such as physical job boards or newspapers, in theory, online advertisement has the potential to reach a larger pool of candidates. In the case of targeted advertising, demographics are an important influencing factor. Here, one potential source of problematic system behaviour is the fact that in many cases demographic data of users (who are the potential targets of the advertising content) is not based on data which they explicitly stated on their user-profile in the given system but a result of predictions from the system. These predictions also have the potential to carry human biases.

The acceptable level of bias needs to be discussed and determined in case of algorithmic systems in recruitment. This can be done on the level of individual systems involved, on the level of any given recruitment process or on a group level. Here it is important to stress that definitions and requirements of individual fairness may not be the same as group fairness. One approach towards reaching a minimum standard of fairness on an individual level is for companies to make sure they do not discriminate against an individual based on protected attributes. The testing of group-level bias is made difficult by the fact that it is very rare to see job adverts that are exactly the same with similar requirements and having the exact same recruitment process as this makes spotting patterns in system behaviour hard. Another element that makes spotting such biases difficult is the complexity of the decision-making process in hiring, where the human decision makers can also be influenced by several types of biases, individual preferences and thought processes, emotions and even irrational thinking. For these reasons the actual metrics of suspected system-bias, problematic system behaviour or discrimination need to be defined quite narrowly.

A different approach from this is not to set requirements for fairness or unbiasedness beyond the prohibition of discrimination that already exists in law, and instead set requirements for transparency where the inputs used, input-output relationships, expected system behaviour and decisions are explained and reasoned.

6. Fairness

The law – at least to a certain extent – should have a requirement as to what fairness definition to use, it should not be left completely up to the company. These should be defined as minimum requirements of fairness as it should be possible for companies to achieve a higher level. It is important to separate law and ethics in the discussion about fairness requirements and also to acknowledge that – at least within one jurisdiction – law is a strict system with set rules whereas ethics can be interpreted and defined differently depending on the values of an individual, group or a society. Here, the legal regulations could serve as minimum standards and ethics could provide guidance for higher levels of fairness and equality standards. One downside of such a system could be that minimum standards may not be enough to inspire developers, vendors, recruiters or employers to try to achieve higher levels and all parties may be aiming for satisfying the mandatory requirements but nothing more. On the other hand, in a system like this, higher standards could be a very important point of marketing (much like the 'fair trade factor'). It is also possible to introduce different levels of compliance where the algorithmic system could get certified for a lower of a higher-level of compliance. Another way to handle relatively low minimum-standards is to gradually increase them with time passing, similarly to sustainability assessments.

7. Incentivising firms to comply with ethics recommendations and to aim for more fair and equitable practices in hiring

A. Creating a repository of what works when it comes to fair and unbiased systems, system evaluations and processes and incentivising relevant stakeholders to be proactive in creating a fair and unbiased recruitment process that is based on equity and inclusion.

B. If investors or institutional investors required ESG factors.

C. Reporting of failures in a 'no blame' culture where it is understood that the advantage that could come from a higher level of failure detection is larger than the advantage that would come from identifying and blaming a particular actor for their mistake/wrongdoing. Here it is important to define what qualifies as a 'failure case'.

8. In conclusion: what aspects could be regulated or at the very least recommended on the level of best practices or guidelines?

A. The prohibition of discrimination (including the ban on using protected characteristics as bases for decisions, unless there is good and justified reason for that)
   a. Developer companies, vendors, recruiters or employers needing to provide evidence that these attributes are not pulled from external sources or candidate profiles and are not being used through proxy characteristics either

B. Transparency of the recruitment process (i.e. mandatory requirement to disclose if algorithmic systems are used in the process, in which stages, for what purpose and which systems)

C. Transparency of systems (including training data, testing methods, inputs, input-output relationships, fairness definitions, system explanations and detailed reasonings of decisions)

D. Provide clear reasoning about the use of specific demographic, skill-based or other criteria in the context of advertisement, assessment or screening

E. Providing statistics on how the system is run, making records on outcomes and keeping them – this is in order to spot group-level bias and to have records in case of individual complaints

F. Providing evidence of advertising claims

G. Assigning responsibilities to developers, vendors, recruiters and employers regarding transparency and ensuring non-discriminatory, fair results

H. Defining specific requirements which are measurable

9. Key implications for the design of a watchdog

A. Unofficial bodies and expert groups that carry out watchdog activities can face severe obstacles when it comes to access to the inside of systems.

B. When carrying out a specific watchdog activity (e.g. investigation), it makes sense to narrow down the scope of the activity to a specific system that carries out a specific type of task in a given stage of the hiring process or to specific types of use-cases.

C. System explanations should not only cover technical details but also how these are interpreted in the sociocultural context of the use.

D. System explanations should be provided on why each criterion was used in the decision.

E. Assessments of systems should take differences in sectors and industries into account.

F. There should be verifications for explanations and marketing promises (this is a potential duty for a watchdog).

G. More research is needed on whether algorithms used in recruitment should be tested before or after deployment.
H. The acceptable level of bias needs to be discussed in case of individual algorithmic systems that are used in recruitment.
I. There are different approaches to overseeing algorithms in recruitment, such as testing for bias, for fairness, for compliance with anti-discrimination legislation or for compliance with transparency standards.
J. It may be difficult to set fairness standards in legislation as there are many fairness definitions both on group and individual level which are debated, sometimes contradictory and very context-dependent in nature.

Appendix 4: Algorithm Watchdog - A CyCAT White Paper

Algorithm Watchdog

A CyCAT White Paper

September 2021
http://www.cycat.io/

Contributors

Michael Rovatsos

Lena Podoletz

Veronika Bogina

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 810105.

Table of Contents

1. Executive summary

This white paper outlines our vision of an algorithmic watchdog - an independent, non-profit entity that provides a public service for investigations of citizens' complaints regarding harms caused by algorithmic systems, operates a public portal to track such disputes and their resolution, offers expert policy advice, publishes recommendations and guidelines based on their investigations and case studies, and builds a library of past cases to document and advance public debate regarding the societal implications of algorithms.

2. Introduction

In recent years, we have seen a rapid increase in the use of products and services that apply algorithmic decision-making processes in the public and private sector. The speed of the evolution of new technologies makes it almost impossible for generic and sectoral legal regulations to keep up with new developments. Many national and international private and public organisations have started to recognise and acknowledge the potential problems that systems using algorithmic decision-making processes might cause to individual citizens, groups, and society as a whole. Many of these organisations have released statements, guidelines and recommendations on what they consider the most serious concerns and issues in the field of algorithmic decision-making. These include issues around transparency, fairness, bias, discrimination, explainability, accountability, appealability of the decisions, privacy and other rights.[32]

The potentially global scale of the problem and the breadth of application areas where issues may arise make it very difficult for these efforts to provide concrete solutions. Therefore, the recommendations articulated in most of these papers remain on the level of general requirements for ethical and trustworthy technologies. Despite these difficulties, provisional solutions to the core issues are being drawn up, for example by the EU's High-Level Expert Group on Artificial Intelligence Policy and Investment Recommendations on Trustworthy AI[33], AccessNow's Human Rights in the Age of Artificial Intelligence[34], AI4People's Ethical Framework for a Good AI Society[35] or The Law Society of England and Wales 2019 Report on Algorithms in the Criminal Justice System[36]. The proposed solutions cover key issues like lawfulness, accountability, transparency, liability, sustainability, and contain guidelines and more concrete recommendations regarding ethical/trustworthy AI and the regulation of algorithmic decision making. Some organisations have also compiled lists and assessments of existing and suggested regulatory solutions regarding artificial intelligence, such as AccessNow's Mapping Regulatory Proposals for Artificial Intelligence in Europe[37] and Algo:Aware's State of the Art Report[38]. It seems, however, that recommendations regarding the oversight and monitoring of algorithmic decision-making systems are missing from the landscape.

[32] Access Now (2018) Human Rights in the Age of Artificial Intelligence; AI Now Institute (2018) AI Now Report; Algo:aware (2018) State of the Art Report: Algorithmic Decision-Making; Muller C (2020) The Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law (Council of Europe Ad Hoc Committee on Artificial Intelligence); European Commission (2020) White Paper: On Artificial Intelligence - A European approach to excellence and trust

[33] EU High-Level Expert Group on AI (2019) Policy and Investment Recommendations on Trustworthy AI

[34] Access Now (2018) Human Rights in the Age of Artificial Intelligence

[35] AI4People (2018) An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations

[36] The Law Society (2019) Algorithms in the Criminal Justice System. The Law Society of England and Wales.

[37] Access Now (2018) Mapping Regulatory Proposals for Artificial Intelligence in Europe

[38] Algo:aware (2018) State of the Art Report: Algorithmic Decision-Making

Following the spirit of these guidelines and documents, we suggest that algorithmic decision-making systems require some form of oversight in order to avoid and mitigate unintended consequences such as discrimination, the use of unfair and biased systems, and in order to provide a higher level of transparency, accountability and possibility to appeal against algorithmic decisions. We recommend that an **Algorithm Watchdog** be able to provide the necessary oversight and other important functions in order to mitigate against the risks of algorithmic decision-making systems and ensure they are deployed to the benefit of society and the economy. Our vision of a working Algorithm Watchdog consists of establishing an independent nonprofit body that will monitor claims from citizens and organisations regarding alleged harms of real-world, deployed products and services that rely on algorithmic components.

3. The need for an Algorithm Watchdog

Watchdog-type activities can be performed in many ways. Algorithms are present in various areas of life and thus their activities and impacts on people, groups and society are highly diverse. This means that in order to evaluate the fairness, bias, discriminative nature, possible harms and consequences of algorithmic decisions, one would need expert knowledge not only on the specific algorithmic system in question, but also on the specific *domain* in which the tool is deployed and on the *context* in which their decisions are made.

For this reason, we believe that an algorithm watchdog should purely consider the algorithmic decision-making systems that are involved in the relevant process. It should only look at decisions purely made by humans without the involvement of algorithmic decision-making systems as long as they serve as a possible explanation for the problematic system outputs. For instance, if there is a complaint from an employer that the candidate recruitment system continuously gives them an output of very similar candidates (e.g. 40-50 year old, white male candidates), the watchdog may decide to ask for more information on input data in order to exclude this as a potential source of bias. Alternatively, it may require domain and legal expert involvement for specific investigations in a particular case.

We believe that such a watchdog may help stakeholders better represent their interests and create some transparency regarding the potential problems caused by algorithmic systems. The overall objective of a watchdog should be to allow citizens to express their concerns regarding algorithmic systems that caused harm, ensure that such claims are addressed by relevant people (legal/domain experts), and that actions are taken. This might involve publicising such disputes and their outcomes, and providing recommendations and guidelines for developers, employers, and systems operators who use these tools as end users. For example, recommendations and guidelines for de-biasing systems, avoiding bias when creating input data, being aware of potential system biases and being able to recognise them could, in the long run, assist in both decreasing algorithmic bias and the influence of human biases on the hiring process.

Based on an initial desk research on relevant regulation and existing authorities and organisations we developed an approach towards watchdog creation. An algorithm watchdog can oversee a set of different factors which can determine the scope of its activities, such as discrimination, bias, fairness, and transparency. However, the list of domains, as well as the list of concerns/harms can be extended, for example to physical, reputational, financial, emotional, psychological harms and human rights. It is also important to provide a platform/service where complaints can be triaged and investigated, without the need of court involvement; that should strengthen citizens' engagement in the process and their ability to make a change.

4. Proposal

The aim is to establish an independent, non-profit entity that provides a transparent service for complaint investigation, operates a public portal to track such disputes and their resolution, offers expert policy advice, publishes recommendations and guidelines based on their investigations and maintains a case library to document past cases and inform best practice.

In the following, we describe a possible structure and set of functions and services an algorithm watchdog should provide (see Figure 1) and the main activities it should carry out (Figure 2).

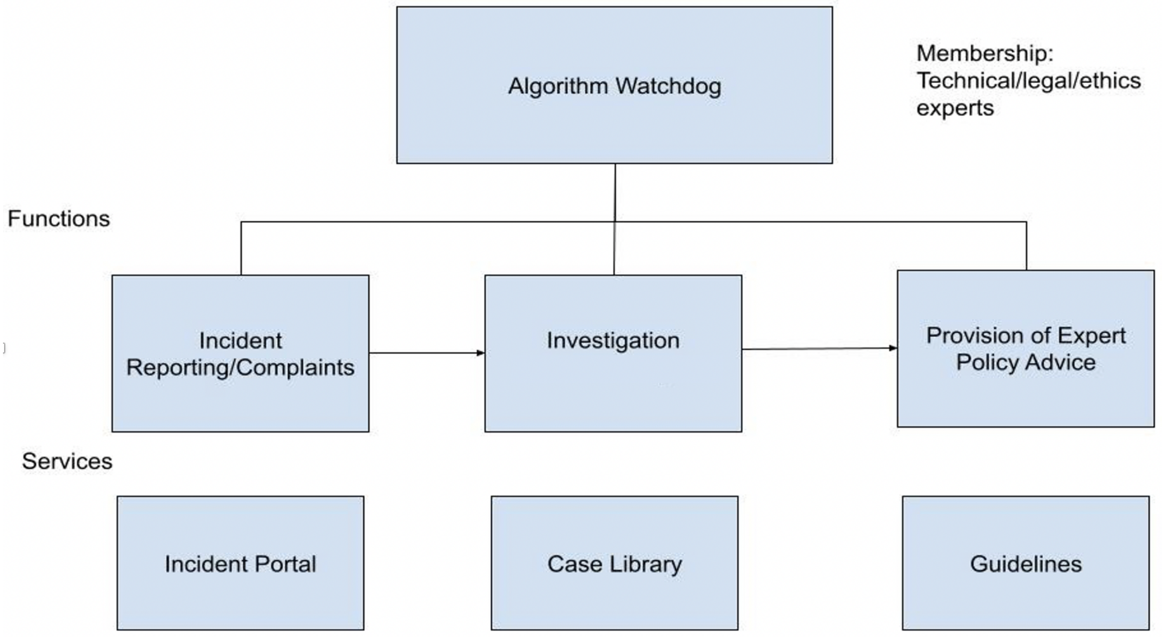
Algorithm Watchdog
Membership: Technical/legal/ethics experts
Functions
Incident Reporting/Complaints
Investigation
Provision of Expert Policy Advice
Services
Incident Portal
Case Library
Guidelines

Figure 1. Algorithm Watchdog functions and services

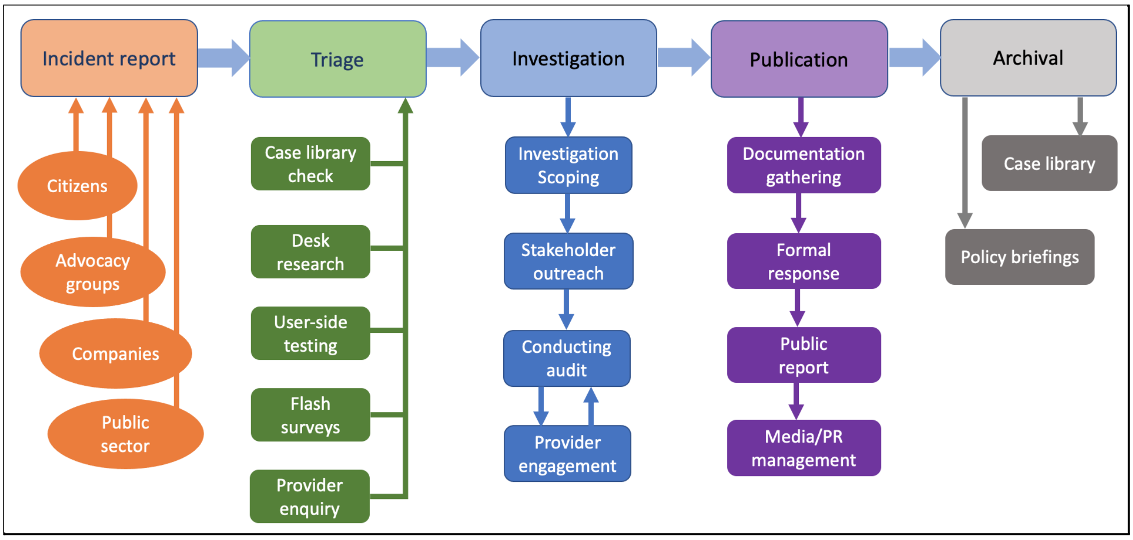

Figure 2. Algorithm Watchdog workflow and connected services

As shown in Figure 1, the functions of an algorithm watchdog would focus on receiving and handling incident complaints, conducting investigations or research projects and providing expert policy advice based on the findings. In addition to these three functions, the algorithm watchdog also provides the following services: incident portal, case libraries and guidelines. The case library represents a database of best practices, which - together with the previous complaints and guidelines created and overseen by an expert group - may serve as both a guideline for lawful and ethical practices as well as an incentive to act as a potential 'best practice' as such commendations can have a positive effect on marketability for both employers and vendors of systems. Below, we will describe a proposed workflow together with the connected services as shown in Figure 2.

a) Incident reporting and complaints

One recommended watchdog function is providing a platform for incident reporting and complaints regarding algorithmic systems. This could provide an opportunity for citizens, advocacy groups, companies as well as representatives of the public sector to submit a report or complaint about a system they believe to be problematic. Overall, the purpose of such Incident Portal is to serve as a basis for investigations and research projects as this would allow both individuals and groups to directly contact the Watchdog. Regarding this step an important policy decision is to decide the threshold for launching an investigation. For instance, would the investigation automatically follow a complaint or should there be some minimum requirements the incident report should cover (e.g., alleged discrimination or a suspected large number of incidents).

b) Triage

Once an incident report had been filed and deemed requiring further consideration, in the triage phase the watchdog would explore the matter through an initial scoping of the problem. Here different methodologies could be used, including:

- checking the case library for similar incidents and applicable cases;
- conducting desk research on the system, the field of application and if necessary, relevant legal regulations;
- carrying out user-side testing in order to gather insight on the system;
- conducting flash surveys amongst users of the system;
- submitting an enquiry to the system provider regarding the incident and the concrete details of the complaint.

Essentially, the triage phase would be used to gather insight on the problem outlined in the complaint, to try and scope the wider context the system is embedded in, and to determine whether an investigation is necessary and feasible.

c) Investigation

The investigation focuses on *investigation and verification of complaints* submitted to the organisation via an Incident Portal platform through the following steps:

- Scoping the problem: Conducting desk research on the specific domain to explore the known issues in the area as well as the system in question. Engaging with existing research is necessary in order to gain a deeper, more structured understanding of the key insights derived from the previous step and to obtain the scientific insight necessary to translate the insights into problems and requirements.

- Stakeholder engagement: Identify relevant stakeholders and gather insight from them on their experiences and the challenges they face when using the system in question. Potential research methods include interviews, focus groups or questionnaires depending on the scale and depth of insights the watchdog is looking to gather.
- Conducting audits: Carry out an audit of the system using technical and social-scientific methods. The details of this step are based on the system in question as well as on whether the system provider grants access to the system and if yes, to what extent.
- Provider engagement: Ideally, the system provider would be contacted during the triage phase and would be in the loop with the investigation making the process cooperative and voluntary. However, one key obstacle in the investigation of systems that unofficial watchdogs without formal powers will most likely face, is the access to the "insides" of the systems. This aspect is a crucial element in determining how the audit is conducted. Without access to the internal structure of the system, the checks are limited to black-box testing methodologies.

d) Publication

After finishing the investigation the watchdog would prepare the acquired documentation for publication. Published materials would include a formal response to the incident report, a full public report and potential media and PR engagement.

e) Archival: case library and guidelines

As mentioned above, one key service provided by the watchdog would be a case library. This would consist of details of past incident reports, the materials gathered throughout the triage and investigation processes as well as final reports and publications. The case library would then feed into future investigations.

The proposed Algorithm Watchdog has the potential to provide expert policy advice on algorithmic decision-making systems, and can help identify the need for new regulation or review of existing proposals, but also inform the technical interpretation and real-life applicability of regulation. In the light of the proposal for the 'Artificial Intelligence Act'[39] in the European Union, such a watchdog can provide guidance on the classification of 'high-risk' and 'low-risk' systems, recommendations for systems to include in the list of 'prohibited AI practices' and assist with the definition of 'conformity assessments'. Both complaint investigations and real-life problem examination can serve as a basis for recommendations and guidelines, and for developing novel technical or regulatory solutions. In addition to this, the knowledge gained from investigations regarding problematic uses of systems and best practices can be translated into a set of recommendations or guidelines.

5. Trialling the investigation method

To assess the feasibility of this concept, we have partially modelled a theoretical investigation. We used this to refine our proposal and to draw recommendations for future work and for the realisation of a real-life watchdog organisation. Without an existing incident reporting platform, we focused on

[39] Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts

recruitment as a high-impact area[40] where potentially biased algorithmic decisions captured the attention of the general public and experts alike. Such decisions may influence the whole of the hiring process and consequently the lives of many jobseekers. We have modelled the following steps: conducting desk research, scoping the problem, engaging with stakeholders, developing recommendations for guidelines.

5.1. Conducting desk research and scoping the problem

As a first step we conducted desk research to scope the field and assess the existing findings and research literature. We found that recruitment decisions not only have an impact on the individual personally affected by the decision (i.e. seeing a job advertisement, getting hired or being promoted) but also more widely across society by counteracting or reinforcing existing biases, creating new ones and impacting social inclusion and equal opportunities. Even though a recent survey suggested that AI-based tools are only used by 5-8% of HR professionals,[41] it is predicted that they will have a far higher use in the future.[42] However, online advertisement through social media, search engines and online job portals[43] is far more frequent, and all of these involve substantial elements of data-driven algorithmic decision making. Despite this, it appears that some candidates have negative sentiments towards the use of algorithms in recruitment, especially if there is no human oversight in the process.[44]

The potential benefits of using algorithmic decision-making tools in recruitment include automating certain high-volume tasks, being able to communicate effectively with prospective candidates, reducing human bias in hiring, and providing a better recruiter, employer, and candidate experience in the hiring process.[45] In principle, these functions can save time and resources, make the hiring process more efficient, and have the potential to allow human resources personnel to focus on less mundane tasks.[46] If these systems are created, trained, tested and deployed correctly, they could have a positive impact on the job market by addressing problems caused by human biases for a long time.

Bias is an issue that has been researched in the field of recruitment for a long time. It has been shown that employers and recruiters can portray bias against people who are of a perceived ethnicity

[40] Upturn (2018) Help Wanted: An Examination of Hiring Algorithms, Equity and Bias. (https://www.upturn.org/reports/2018/hiring-algorithms/); AI Now Institute (2018) AI Now Report (https://ainowinstitute.org/AI_Now_2018_Report.pdf); https://www.viasto.com/en/blog/survey-artificial-intelligence-in-hr/

[41] https://www.oracle.com/a/ocom/docs/artificial-intelligence-in-talent-acquisition.pdf

[42] https://www.viasto.com/en/blog/survey-artificial-intelligence-in-hr/

[43] For instance, a survey showed that 52% of jobseekers prefer the use of online job-seeking platforms to traditional methods. https://www.glassdoor.co.uk/employers/resources/40-hr-and-recruiting-stats-for-2020/

[44] https://www.kornferry.com/about-us/press/putting-ai-in-place-artificial-intelligence-should-be-part-of-the-recruiting-process-but-cant-re
place-the-human-touch-korn-ferry-survey, https://www.cfsearch.com/wp-content/uploads/2019/10/James-Wright-The-impact-of-artificial-intelligence-within-the-recruitment-industry-Defining-a-new-way-of-recruiting.pdf, https://markets.businessinsider.com/news/stocks/nearly-9-in-10-americans-would-feel-uncomfortable-with-an-artificial-intelligence-job-interview-app-being-used-to-screen-candidates-1028737898

[45] https://techrseries.com/others/ai-in-recruitment-benefits-and-challenges/, https://theundercoverrecruiter.com/benefits-ai-recruitment/

[46] https://www.hrmorning.com/articles/artificial-intelligence-technology/

[47], conventionally less attractive[48], have a particular accent[49] or have studied or worked in a foreign country[50]. Some studies also present evidence to the contrary in particular research settings, such as in case of gender[51] or in case of race and gender[52]. The importance of cultural differences has also been shown, for example, in the context of evaluating job applicant behaviour, such as excited and calm states in interview settings[53]. It seems that algorithmic systems used in the hiring funnel are not free from biases either[54]. For instance online advertisements have been shown to be biased in terms of gender and ethnicity.[55] Regardless of their origin, biases in hiring practices can severely impact jobseekers by limiting the "employment opportunities for historically excluded groups".[56]

5.2. Stakeholder engagement

In a second step, we conducted six semi-structured interviews with stakeholders. The purpose of these interviews was to gather insight on the challenges algorithmic systems used in recruitment present as well as to understand the landscape from the perspective of the people who work in the field of recruitment or have been involved with the field from a policy-making angle, such as representatives of recruiters, large-scale employers and advocacy groups. Below we present the key unresolved issues stakeholders are not able to address given their current knowledge of these systems and the algorithms that underpin them:

1. Lack (or very low levels) of transparency when it comes to:
    a. The development process (e.g. was the development team diverse, what fairness and other metrics were applied, where training data came from, how testing was conducted)
    b. The internal mechanics of the tool (e.g. what data it uses, how it weighs different variables)
    c. Potential system biases
2. Biases in algorithmic systems

[47] Benedick Jr. M et al (1992) Discrimination against Latino Job Applicants: A Controlled Experiment. *Human Resource Management* 30(4):469-484; Bertrand M et al (2004) Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Dicrimination. *The American Economic Review* 94(4):991-1013; Watson S et al (2011) The Effect of Name on Pre-Interview Impressions and Occupational Stereotypes: The Case of Black Sales Job Applicants. *Journal of Applied Psychology* 41(10); Cotton J et al (2008) "The Name Game": Affective and Hiring Reactions to First Names. *Journal of Managerial Psychology* 23(1):18-39; Jacquemet N et al (2013) Indiscriminate discrimination: A correspondence test for ethnic homophily in the Chicago labor market. *Labour Economics* 2012(19):824-832; Kang SK et al (2016) Whitened Résumés: Race and Self-Preservation in the Labor Market. *Administrative Science Quarterly* 61(3):469-502

[48] Bóo FL et al (2013) The labor market return to an attractive face. *Economics Letters* 2013(118):170-172

[49] Segrest S et al (2006) Implicit sources of bias in employment interview judgements and decisions. *Faculty Publications: Scholarly Works*, University of South Florida St. Petersburg

[50] Oreopoulos P (2009) Why do skilled immigrants struggle in the labor market? A field experiment with six thousand resumes. *NBER Working Paper Series*. Cambridge, MA: National Bureau of Economic Research

[51] Smith FL et al (2007) The Name Game: Employability Evaluations of Prototypical Applicants with Stereotypical Feminine and Masculine First Names. *Sex Roles* 52(1/2):63-82

[52] Darolia R et al (2016) Race and gender effects on employer interest in job applicants: new evidence from a resume field experiment. *Applied Economics Letters* 23(12):853-856

[53] Bencharit LZ et al (2019) Should Job Applicants Be Excited or Calm? The Role of Culture and Ideal Affect in Employment Settings. *Emotion* 19(3):377-401

[54] Upturn (2018) Help Wanted: An Examination of Hiring Algorithms, Equity and Bias. (https://www.upturn.org/reports/2018/hiring-algorithms/)

[55] Datta A et al (2015) *Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice and Discrimination*; Ali M et al (2019) *Discrimination through optinization: How Facebook's ad delivery can lead to skewed outcomes*; Lambrecht A et al (2019) Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Management Science* 65(7):2966-2981

[56] Benedick Jr. M et al (1992) Discrimination against Latino Job Applicants: A Controlled Experiment. *Human Resource Management* 30(4):469-484. p238

   a. How does this compare to biases in human decisions
   b. How to detect algorithmic bias in recruitment decisions
3. Limitations of currently available tools (e.g., the set of variables they consider are very limited compared to what a human recruiter checks)
4. No incentives for vendors to disclose information related to the above concerns
5. Limitations of relevant knowledge and skills of users and subjects to evaluate any information related to the above concerns, even if it is disclosed
6. Lack of accountability for biased decisions when it comes to developers and vendors
7. Lack of clear guidelines, standards or clear interpretations of current regulations of algorithmic decision-making tools used in the field of recruitment

5.3. Proposing guidelines for the issues discovered through the investigation

In order to model the last step of turning the results of an investigation into recommendations for guidelines, we have organised two expert workshops where we presented our findings with the aim of gathering understanding of the state-of-the-art knowledge when it comes to the key issues mentioned by our interviewees. This group included academics and industry representatives with either technical, legal, or domain expertise, and was tasked with producing recommendations that could inform future regulation, guidelines, and codes of conduct for the industry related to the problems highlighted by our interview participants.

Key recommendations regarding the area of recruitment emerging from this discussion are as follows:

1. The prohibition of discrimination (including the ban on making decisions on the basis of protected characteristics, unless there is good reason for doing so) implies that product vendors, recruiters, and employers should be expected to provide evidence that these attributes are not pulled from external sources or candidate profiles and are not influencing decisions through proxy characteristics.
2. It is reasonable to expect transparency in terms of the recruitment process (i.e. a mandatory requirement to disclose which, if any, algorithmic systems are used in the process, in which stages, and for what purpose), and in terms of the algorithmic components used (including training data, testing methods, inputs, input-output relationships, fairness definitions, available explanations, and detailed reasoning behind recommendations and decisions).
3. Clear justifications should be provided when specific demographic, skill-based or other criteria are applied during advertisement, assessment or screening. Advertising claims need to be backed by evidence from the system's behaviour, and clear requirements need to be defined against which all of the above can be measured.
4. Organisations using the systems should be expected to be able to provide documentation on how the system is run, creating and keeping records on previous outcomes including statistical analysis of outcomes in order to spot group-level bias, as well as to keep records of individual complaints.
5. The responsibilities regarding ensuring non-discriminatory and fair results and transparency described above in the different steps of the recruitment process need to be assigned to developers, vendors, recruiters and employers for every specific use of a system.

5.4. Conclusions from the 'trial investigation'

Unsurprisingly, the experience of working through this case study clearly showed that there are gaps in the oversight and management of risks that might emerge from new algorithmic systems deployed in real-world domains, and that any recommendations that might emerge from such expert groupings must trigger further steps to influence industry and policy makers. But what our work has also revealed is that it is possible to establish a productive exchange with a range of stakeholders to capture their concerns and feed them into an expert body that can provide authoritative advice on specific measures to be taken in order to mitigate the societal risks created by algorithms in certain application domains.

The expert workshops also presented us with some key points and insights when it comes to the workings of a potential Algorithm Watchdog. The following is a list of the key implications derived from our work for the design of an algorithm watchdog organisation:

1. Unofficial bodies and expert groups that carry out watchdog activities can face severe obstacles when it comes to access to the inside of systems.
2. When carrying out a specific watchdog activity (e.g. investigation), it makes sense to narrow down the scope of the activity to a specific system that carries out a specific type of task in a given stage of the hiring process or to specific types of use-cases.
3. System explanations should not only cover technical details but also how these are interpreted in the sociocultural context of the use
4. Assessments of systems should take differences in sectors and industries into account.
5. Explanations and marketing promises given for these systems should be verified.
6. More research is needed on whether and how algorithms used in a given context should be tested before or after deployment.
7. The acceptable level of bias needs to be determined for individual algorithmic systems that are used in a specific application context.
8. There are different approaches to overseeing algorithms, such as testing for bias, for fairness, for compliance with anti-discrimination legislation, or for compliance with transparency standards.
9. It may be difficult to set fairness standards in legislation as there are many fairness definitions both on group and individual levels which are debated, sometimes contradictory and very context-dependent in nature.

Finally, to validate the proposed structure for the Algorithm watchdog we held an interdisciplinary online event with the participation of experts on AI regulation and representatives from organisations that carry out watchdog-type activities. This event shed light on some of the questions that require further research. These are:

a) How can algorithm auditing methodologies be tested and validated?
b) What type of training do algorithm auditors need?
c) Whose interests will watchdogs serve and how can watchdogs deal with competing values within society?
d) What is the best and safest way to ensure access to an algorithmic system and all the relevant data for auditing?
e) If a watchdog organisation does not hold official power, what can be the effect of an audit?
f) If a watchdog organisation holds official power, what would be the effective remedies?
g) What level of citizen involvement is needed and what is the best way to ensure that engagement?
h) How can a non-profit, independent watchdog remain sustainable?

h) How to overcome the limitations presented by GDPR when it comes to algorithm auditing in an EU environment?
i) How can watchdogs demonstrate trustworthiness?

7. Conclusions

This paper has presented a prototypical concept for an Algorithm Watchdog, conceptualised as an neutral expert body that can mediate between the publics of citizens that are potentially affected by algorithmic systems, the organisations developing and deploying them, and regulators and policy makers. The main purpose of such a watchdog would be to enable a rigorous investigation of concerns expressed by citizens through a publicly available complaints platform, enable a transparent and moderated dialogue between users and providers of algorithmic systems with potential substantial societal impact, mobilise multidisciplinary expertise to articulate recommendations that inform best practice and future regulation, and conduct the research required to develop these recommendations.

While our initial work has demonstrated the feasibility and usefulness of some of the core elements of this proposal, it has naturally been limited to an in vitro emulation of some of the proposed watchdog activities, and further research will be needed to examine how such an entity might be established.