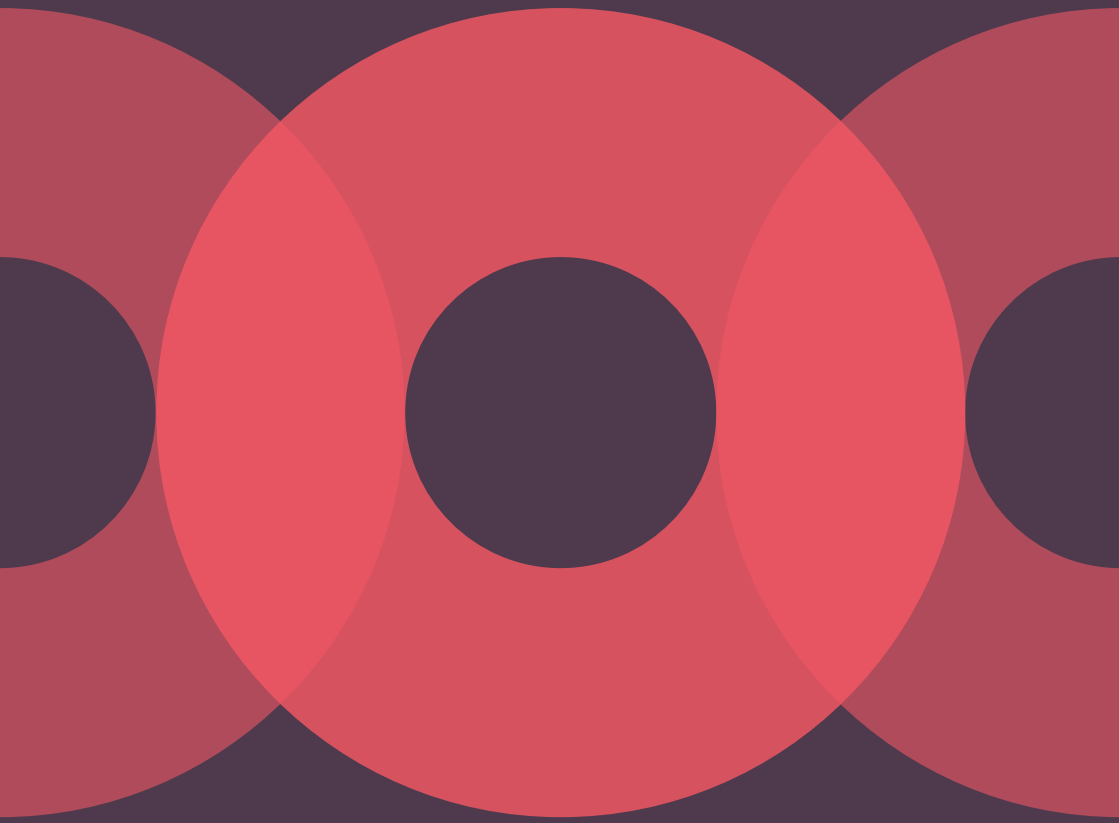
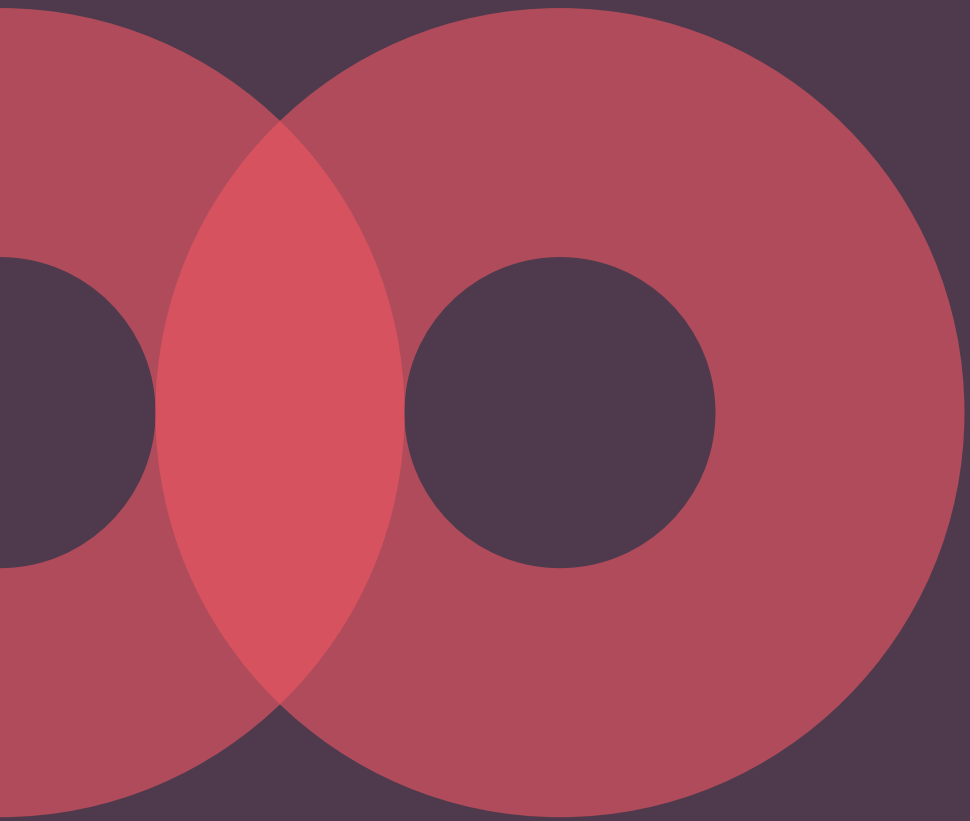
Diskrimineringsombudsmannen

Rapport 2023:6

AI och risker för diskriminering i arbetslivet





AI och risker för diskriminering i arbetslivet

Denna pdf är tillgänglig. Andra alternativa format, exempelvis Daisy, lättläst eller punktskrift, kan beställas vid behov. Kontakta då DO på order@do.se.

Diskrimineringsombudsmannen (DO)

R27 2023 PDF
ISBN 978-91-88175-48-9

Illustration: Stina Wirsén.
Layout: Kollijox AB.

Förord

Användandet av artificiell intelligens (AI) och annat automatiserat beslutsfattande har ökat kraftigt i arbetslivet under de senaste decennierna. Tillämpningen av sådana tekniska lösningar kan innebära risker för diskriminering och utgöra hinder för lika rättigheter och möjligheter på flera plan.

Risker för diskriminering kan uppkomma eftersom den data som AI-system tränas på återspeglar historiska och befintliga ojämlikheter och grupprelaterade skillnader i arbetslivet. Om ett begränsat antal AI-system, som reproducerar dessa skillnader, dominerar inom ett område kan det i sin tur leda till konsekvent och systematisk diskriminering av ett stort antal personer. Brist på insyn och transparens är en annan utmaning som kan öka risken för diskriminering och försvåra tillsynen av systemen.

DO ser allvarligt på de risker för diskriminering som användandet av AI och automatiserat beslutsfattande kan medföra. Utvecklingen av AI och annat automatiserat beslutsfattande är därför ett område där DO anser att det krävs ett aktivt och medvetet arbete av både AI-utvecklare och AI-användare, arbetsgivare, för att motverka diskriminering. Det är även ett område som DO följer framöver för att förebygga risker för diskriminering.

Följande rapport är en redovisning av ett uppdrag som regeringen gett till DO. I uppdraget ingår att fördjupa kunskapen om risker för diskriminering vid användandet av AI och annat automatiserat beslutsfattande inom arbetslivet. I arbetet med regeringsuppdraget har DO anlitat två forskargrupper. Den ena forskargruppen har genomfört en systematisk forskningsöversikt för att kartlägga vilka risker för diskriminering som identifierats i forskning vid användning av AI och annat automatiserat beslutsfattande.[1] Den andra forskargruppen fick i uppgift att undersöka i vilken utsträckning och i vilka sammanhang som arbetsgivare använder sig av AI och annat automatiserat beslutsfattande och vilken kunskapsnivå de har om risker för diskriminering.[2]

Förhoppningen är att rapporten ska ge ökad kunskap om risker för diskriminering när AI och annat automatiserat beslutsfattande används inom arbetslivet. DO vill även att arbetsgivare blir mer medvetna om sin användning av sådana tekniska lösningar och att de tar ett ökat ansvar för att motverka diskriminering genom att följa diskrimineringslagen.

Solna den 15 november 2023

Lars Arrhenius
Diskrimineringsombudsman

1 Sima Nourali Wolgast och Martin Wolgast, Institutionen för psykologi på Lunds universitet.
2 Stefan Larsson och James White, Institutionen för teknik och samhälle, Lunds Tekniska Högskola, Lunds universitet.

Innehåll

Sammanfattning

Denna rapport är en slutredovisning av regeringens uppdrag till DO att fördjupa kunskapen om risker för diskriminering vid användandet av artificiell intelligens (AI) och annat automatiserat beslutsfattande inom arbetslivet.[3]

Syfte

I uppdraget ingår att kartlägga vilka risker för diskriminering som användandet av AI och annat automatiserat beslutsfattande kan medföra, och i vilken utsträckning och i vilka sammanhang som arbetsgivare använder sådana tekniska lösningar. Vidare ingår det i uppdraget att belysa kunskapsnivån hos arbetsgivare om förhållanden som kan leda till diskriminering när det gäller AI och annat automatiserat beslutsfattande.

Metod och tillvägagångsätt

DO har anlitat två forskargrupper i arbetet med regeringsuppdraget. Den ena forskargruppen har genomfört en systematisk forskningsöversikt för att kartlägga vilka risker för diskriminering som har identifierats i forskning vid användning av AI och annat automatiserat beslutsfattande.[4] Den andra forskargruppen fick i uppgift att undersöka i vilken utsträckning och i vilka sammanhang som arbetsgivare använder sig av AI och annat automatiserat beslutsfattande och vilken kunskapsnivå arbetsgivare har om risker för diskriminering.[5]

3 Arbetsmarknadsdepartementet (2022), Regeringsbeslut 2022-06-07, A2022/00877.
4 Sima Nourali Wolgast och Martin Wolgast, Institutionen för psykologi på Lunds universitet.
5 Stefan Larsson och James White, Institutionen för teknik och samhälle, Lunds Tekniska Högskola, Lunds universitet.

De forskargrupper som DO har anlitat har i delar använt en annan definition av diskriminering än diskrimineringslagens definition.

Definition – AI är ett svårfångat begrepp

De forskare som DO har anlitat har även valt att använda något olika definitioner av AI och annat automatiserat beslutsfattande. Det i sig har inga större konsekvenser för de slutsatser som kan dras. I stället speglar det svårigheterna med att kartlägga ett område där själva definitionen är undanglidande och i ständig rörelse i takt med teknisk utveckling och tillämpning.

Resultat

AI och annat automatiserat beslutsfattande används i ökande omfattning av arbetsgivare i dag, även om dessa inte alltid är medvetna om vilken teknik de använder. Detta kan DO övergripande konstatera, med utgångspunkt i de två undersökningar som vi har beställt. Likaså kan vi konstatera att AI och annat automatiserat beslutsfattande medför risker för diskriminering och att arbetsgivare i dag endast till viss del är medvetna om dessa risker. Det innebär att kunskapsnivån behöver höjas både om när arbetsgivare använder AI och annat automatiserat beslutsfattande och hur risker för diskriminering kan förebyggas.

Rapporten kan läsas som en antologi med fristående kapitel.

Rapportens sammanfattande resultat

Motstridiga forskningsresultat och forskningsluckor

Det finns stora forskningsluckor om risker för diskriminering vid användandet av AI och annat automatiserat beslutsfattande i arbetslivet. Det finns till exempel behov av mer forskning och kunskap om de digitala plattformarnas roll inom svensk rekryteringspraxis och deras hantering av risker för diskriminering i relation till AI och annat automatiserat

beslutsfattande. Detta trots att det sannolikt kommer att variera i vilken utsträckning digitala plattformtjänster använder automatisering och AI-baserade lösningar som beslutsstöd.

Forskningen pekar i flera fall åt olika håll. Vissa forskningsresultat visar på möjligheter med AI och annat automatiserat beslutsfattande när det gäller att förebygga diskriminering. Samtidigt lyfter andra forskningsresultat fram riskerna för diskriminering när AI och annat automatiserat beslutsfattande används inom samma område och på liknande sätt.

Återspegling av ojämlikheter på arbetsmarknaden

Risker för diskriminering kan uppkomma eftersom den data som AI-system tränas på återspeglar historiska och befintliga ojämlikheter och grupprelaterade skillnader i arbetslivet. Denna grundläggande omständighet riskerar exempelvis att leda till att systemen tolkar denna data som att etablerade mönster av över- och underrepresentation mellan grupper inom olika branscher eller på olika befattningsnivåer återspeglar skillnader i kompetens och lämplighet. Systemen riskerar då att fatta eller föreslå beslut som reproducerar skillnader mellan grupper.

Automation bias – att okritiskt acceptera utfallet av systemen

Mänskliga beslutsfattare riskerar att bli mindre noggranna och kritiska när de använder AI och annat automatiserat beslutsfattande. Det finns en inbyggd risk för så kallad automation bias. Det innebär att beslutsfattare okritiskt accepterar utfallet av systemen som något neutralt, objektivt och värderingsfritt. Risken är att diskriminerande utfall som systemen utvecklar inte upptäcks och korrigeras om användarna inte är medvetna om, eller organisatoriskt tillåts värdera, att dessa system – precis som alla andra metoder för informationsbearbetning och beslutsfattande – har såväl för- som nackdelar och måste granskas kritiskt.

AI-systemen kan leda till systematisk diskriminering

System baserade på AI och annat automatiserat beslutsfattande kan leda till konsekvent och systematisk diskriminering. Det kommer att finnas en stor variation i hur diskriminering på grund av mänskliga beslutsfattare ser ut och vilka konsekvenser denna diskriminering får i enskilda fall, även om systemen skulle vara mindre diskriminerande än mänskliga beslutsfattare i de enskilda fallen. Sådan diskriminering har en föränderlighet som drivs av en rad omständigheter, vilket har med såväl person- som situationsfaktorer att göra. Den diskriminering som sker riskerar emellertid att bli mer enhetlig, konsekvent och systematisk om ett begränsat antal system baserade på AI och annat automatiserat beslutsfattande, eller flera olika system som fungerar på liknande sätt, dominerar inom ett område.

Skillnad mellan upplevd och faktisk användning av AI och annat automatiserat beslutsfattande

Få arbetsgivare är medvetna om i vilken utsträckning de använder sig av AI och annat automatiserat beslutsfattande. Det finns en åtskillnad mellan upplevd och faktisk användning av AI och annat automatiserat beslutsfattande hos arbetsgivare som har deltagit i undersökningen. Arbetsgivare är inte medvetna om att de redan nu använder sig av AI och annat automatiserat beslutsfattande, till exempel vid rekrytering eller i det löpande arbetsledande personalarbetet såsom vid hanteringen av lönesystem med mera. Det innebär att det finns en risk att dessa tekniska lösningar både är väl integrerade i vardagen och ändå ibland är okritiskt och omedvetet betrodda av de personer som använder dem.

Bristande transparens och möjlighet till granskning leder till risker för diskriminering

Stora företag med rekryteringsbehov upphandlar delar till rekryteringsföretag som i sin tur använder sig av tredjepartssystem, det vill säga stora plattformar som ägs av globala teknikbolag. Detta kan försvåra transparens och möjlighet till granskning då de datadrivna digitala

tjänster som automatiserar matchning är komplicerade att granska. Bristande transparens kan i sin tur leda till risker för diskriminering.

Undersökningen pekar på att systemen inte kan undkomma mänskliga fördomar och ojämlikheter, eftersom systemen utvecklas och även tränas av människor och bygger på data som representerar mänskligt språk och samhälle. Det innebär att AI och annat automatiserat beslutsfattande alltid medför risker för diskriminering. AI- systemen behöver därmed granskas och hanteras för att nå de kostnadseffektiva och precisa matchningseffekter som förväntas uppnås vid rekrytering via dessa tredjepartslösningar.

DO:s slutsatser

Utifrån de sammanfattande resultaten vill DO lyfta fram följande slutsatser.

AI-systemen måste bli mer transparenta för att motverka diskriminering

AI-systemen måste bli mer transparenta då bristande transparens gör det svårt att identifiera risker för diskriminering. En tillsynsmyndighet ska kunna granska hur ett AI-system utvecklas. Samtidigt är de flesta större digitala plattformsföretag väldigt svåra att få tillgång till. Det orsakar en särskild utmaning vid tillsyn, där även de storskaliga, ofta globala, tekniska plattformarna behöver kunna granskas utifrån (skalbara) risker för diskriminerande utfall.

Dessutom måste ansvarsfördelningen bli tydligare kring vem som bär ansvar för risker för diskriminering. Arbetsgivaren har ansvar för att ingen utsätts för diskriminering. Men det finns en överhängande risk att varken företagen eller rekryterarna kan kontrollera hur de AI-system som ingår i den digitala matchningen av kandidater via sociala medier eller andra tredjepartstjänster för rekrytering faktiskt säkerställer att de inte

diskriminerar. Utan möjlighet till ansvarsutkrävande riskerar felaktiga eller diskriminerande AI-system att inte upptäckas och åtgärdas.

Det behövs också rättsliga tolkningar i ljuset av AI-fältets utveckling som skulle kunna belysa risker för diskriminering. Arbetet inom EU med att ta fram en förordning om AI är ett steg i riktning mot att reglera vissa AI-system. För att betona att skyddet mot diskriminering gäller parallellt med AI-förordningen har DO påtalat att det i AI-förordningen bör framgå att ett godkännande enligt förordningen inte påverkar leverantörers och användares ansvar för diskriminering som sker genom sådana AI-system.[6]

Risker för systematisk diskriminering och representationsbias i AI-system måste förhindras

Den ökade användningen av AI-system kan bidra till diskriminerande processer och utfall som blir mer systematiska och konsekventa. Risker för systematisk diskriminering måste därför förhindras.

AI-system bygger på data som återspeglar historiska och befintliga ojämlikheter och grupprelaterade skillnader i arbetslivet, vilket kan leda till strukturella fördomar. De har även inbyggda representationsbias, vilket innebär att träningsdatan inte representerar alla grupper. Det påverkar i sin tur AI-systemens precision negativt för dessa grupper. Det finns också ojämlika och diskriminerande strukturer i samhället som reproduceras i exempelvis språk och texter som används som träningsdata. Eftersom AI-modeller tränas på stora datamängder kan man inte utesluta att alla fördomar de har också kommer att vara utbredda. Detta ökar risken för diskriminering i stor omfattning, vilket också kan vara svårt att upptäcka.

För AI-utvecklare och AI-användare, arbetsgivare, blir det därmed viktigt att upptäcka, åtgärda och förhindra sådan så kallad bias i AI-systemen.

6 Diskrimineringsombudsmannen, 2021-06-23.

Arbetsgivare måste öka sin medvetenhet om AI och annat automatiserat beslutsfattande och arbeta med aktiva åtgärder

Att man som arbetsgivare känner till att man använder AI och annat automatiserat beslutsfattande är avgörande för att arbeta förebyggande för att motverka diskriminering vid tillämpning av tekniken.

Resultaten i rapporten visar att den faktiska användningen bland arbetsgivare är högre än den upplevda användningen av sådana tekniska lösningar. Det tycks vara svårt för arbetsgivare att vara helt säkra på när de använder system som bygger på någon form av AI eller automatiserat beslutsfattande. Kunskapen om vad AI och annat automatiserat beslutsfattande är, och medvetenheten om när arbetsgivare använder tekniken, måste därmed öka. I detta sammanhang är det även viktigt att arbetsgivare aktivt arbetar för att motverka automation bias. Det vill säga att tilltron till de automatiserade besluten leder till ett passivt, okritiskt och oreflekterat förhållningssätt hos den beslutande arbetsgivaren.

Kunskapsnivån hos arbetsgivare är inte tillfredsställande när det gäller de risker för diskriminering som finns vid användningen av AI och annat automatiserat beslutsfattande. En annan slutsats är därför att arbetsgivare måste ta ett ökat ansvar för att motverka diskriminering vid användningen av AI och annat automatiserat beslutsfattandegenom att följa diskrimineringslagen. Arbetsgivare har en skyldighet att leva upp till diskrimineringslagens krav på aktiva åtgärder, oavsett vilka tekniska lösningar de använder.

Avslutningsvis kan vi konstatera att utvecklingen av AI och annat automatiserat beslutsfattande är ett område som DO och andra likabehandlingsorgan i Europa behöver följa för att bidra till att förebygga diskriminering.

Summary

This report is a final report for the assignment that DO was given by the Swedish Government to improve our knowledge of risks of discrimination in the use of artificial intelligence (AI) and other types of automated decision-making within working life.[7]

Aim

This assignment includes charting the risks of discrimination that may result from the use of AI and other types of automated decision-making and to what extent and in which contexts employers are using such technical solutions. Furthermore, the assignment includes highlighting the level of knowledge among employers about the conditions that may lead to discrimination in the context of AI and other types of automated decision-making.

Method and approach

DO engaged two groups of researchers for the work entailed in this government assignment. One of the groups conducted a systematic literature review in order to chart the risks of discrimination that have been identified when researching the use of AI and other types of automated decision-making.[8] The other group investigated the extent to which and in what circumstances employers are using AI and other types of automated decision-making and what level of knowledge employers have

7 Ministry of Employment (2022), Government decision 07/06/2022, A2022/00877.
8 Sima Nourali Wolgast and Martin Wolgast, Department of Psychology at Lund University.

about the risks of discrimination.[9] The researcher groups engaged by DO have, in some places, used a different definition of discrimination than that which is used in the Discrimination Act.

Definition – AI is a concept that is not easily defined

The researchers engaged by DO have also chosen to use somewhat differing definitions of AI and other types of automated decision-making. This has no major consequences in terms of the conclusions that can be drawn. Instead it reflects the difficulties in charting an area where the definition itself is elusive and in constant flux as the technology develops and applications change.

Results

AI and other types of automated decision-making are nowadays increasingly being used by employers, although they are not always aware of which technology they are using. This is one overall conclusion DO is able to draw on the basis of the two studies we have commissioned. We are also able to conclude that AI and other types of automated decision-making result in risks of discrimination and that employers are currently only partly aware of these risks. This means that the level of knowledge needs to be raised about both when employers are using AI and other types of automated decision-making, and how discrimination risks can be prevented.

The final report can be read as an anthology with stand-alone chapters.

9 Stefan Larsson and James White, Department of Technology and Society, Faculty of Engineering, Lund University.

The results of the final report in summary

Conflicting research results and research gaps

There are large gaps in the research into risks of discrimination when using AI and other types of automated decision-making in working life. For example, there is a need for more research and knowledge about the role of digital platforms within Swedish recruitment practices and how these are handling risks of discrimination in relation to AI and other types of automated decision-making. This is in spite the fact that there will probably be variation in the extent to which digital platform services use automation and AI-based solutions to support decision-making.

In several cases the research points in different directions. Some research results indicate the potential for AI and other types of automated decision-making when it comes to preventing discrimination. At the same time, other research results highlight the risks of discrimination when AI and other types of automated decision-making are being used within the same area and in a similar manner.

Reflection of inequalities in the labour market

Risks of discrimination may arise due to the data on which AI systems are trained reflecting historical and existing inequalities and group-related differences in working life. For example, there is a risk that this fundamental circumstance will lead to the systems interpreting these data as evidence that established patterns of over and under-representation among groups within different industries or in different ranks reflect differences in competence and suitability. There is then a risk that systems will make or propose decisions that reproduce existing differences between groups.

Automation bias – uncritically accepting the outcome of the systems

There is a risk that human decision makers will becoming less thorough and critical when using AI and other types of automated decision-making. There is an inherent risk of automation bias. This means that decisions

makers uncritically accept the outcome of these systems as something neutral, objective and non-judgemental. The risk is that discriminatory outcomes developed by the systems are not detected and corrected if the users are not aware, or allowed to organisationally evaluate, that these systems – just like all other methods of information processing and decision-making – have disadvantages that must be critically appraised.

AI systems can lead to systematic discrimination

Systems based on AI and other types of automated decision-making may lead to consistent and systematic discrimination. There will be a large variation in different types of discrimination caused by human decision makers and the consequences of this discrimination in individual cases, even if the systems were to be less discriminatory than human decision makers in each individual case. Such discrimination is variable and driven by several circumstances that relate to both individual and situational factors. However, there is a risk that the discrimination which takes place will become more standardised, consistent and systematic if a limited number of systems based on AI and other types of automated decision-making, or several different systems that work in a similar way, are dominant within one area.

The difference between perceived and actual use of AI and other types of automated decision-making

Few employers are aware of the extent to which they are using AI and other types of automated decision-making. There is a difference in perceived and actual use of AI and other types of automated decision-making among employers who have participated in the investigation. Employers are not aware that they are already using AI and other types of automated decision-making, for example, for recruitment or in ongoing HR management work such as administration of payroll systems etc. This means that there is a risk of these technical solutions being both well-integrated into day-to-day work but still sometimes uncritically and unconsciously trusted by the people who use them.

Lack of transparency and audit opportunities leads to risks of discrimination

Large companies with recruitment needs procure aspects of this work from recruitment companies, which in turn utilise third-party systems, i.e. large platforms owned by global tech companies. This may impair transparency and audit opportunities as the data-driven digital services that automate matching are complicated to audit. A lack of transparency may in turn lead to risks of discrimination.

The investigation indicates that the systems are not able to escape human prejudices and inequalities as the systems are developed and trained by people, as well as being based on data that represent human language and society. This means that AI and other types of automated decision-making always entail risks of discrimination. Consequently, AI systems need to be audited and managed with the aim of achieve the cost-effective and precise matching effects that are to be expected when recruiting via these third-party solutions.

DO's conclusions

Based on the summarised results DO wants to highlight the following conclusions.

The AI systems must become more transparent in order to combat discrimination

AI systems must become more transparent as a lack of transparency makes it difficult to identify risks of discrimination. A supervisory authority must be able to audit how an AI system is developed. At the same time, it is very difficult to gain access to most of the larger digital platform companies. This makes supervision especially challenging where it is necessary to be able to audit the large-scale, often global, technical platforms on the basis of (scalable) risks of a discriminatory outcome.

In addition, the division of responsibility must become clearer with regard to who is responsible for risks of discrimination. The employer is responsible for ensuring that no one is subjected to discrimination. However, there is an imminent risk that neither the companies nor the recruiters are able to verify how the AI systems included in the digital matching of candidates via social media or other third-party recruitment services actually ensure these are not discriminatory. Without opportunities to demand accountability there is a risk that faulty or discriminatory AI systems will not be detected and rectified.

There is also a need for legal interpretations in light of developments in the field of AI that could highlight risks of discrimination. The work carried out within the EU to produce the so called AI Act is one step on the path towards regulating certain AI systems. In order to emphasise that the protection against discrimination applies in parallel with the AI regulation, DO has argued that the AI regulation should make it clear that an approval in accordance with the regulation does not affect the liability of suppliers and users for discrimination that takes place through the use of such AI systems.[10]

Risks of systematic discrimination and representation bias in AI systems must be prevented

The increasing use of AI systems may contribute to discriminatory processes and outcomes that become more systematic and consistent. Risks of systematic discrimination must therefore be prevented.

AI systems are built using data that reflect historical and existing inequalities and group-related differences in working life, which may lead to structural prejudices. They also have built-in representation bias, which means that the training data do not represent all groups.

10 Equality Ombudsman, 23/06/2021.

This in turn has a detrimental impact on the precision of AI systems for these groups. There are also unequal and discriminatory structures in society that are reproduced in, for example, language and texts used as training data. As AI models are trained on large data sets it cannot be ruled out that all prejudices present in these will also be widespread. This greatly increases the risk of discrimination, which may also be difficult to detect.

Consequently, it becomes important that AI developers and AI users (employers) detect, rectify and prevent such bias in AI systems.

Employers must increase their awareness of AI and other types of automated decision-making and work with active measures

For an employer, an awareness that AI and other types of automated decision-making are being used is vital to allowing preventive action to be taken to combat discrimination when applying this technology.

The results of the report indicate that actual use by employers is higher than the perceived use of such technical solutions. It appears to be difficult for employers to be completely certain about when they are using systems based on some form of AI or other types of automated decision-making. Knowledge of what AI and other types of automated decision-making are and awareness of when employers are using this technology must therefore be improved. In this context, it is also important that employers work actively to combat automation bias. In other words, the confidence that automated decisions lead to a passive, uncritical and unreflecting approach from the employer making these decisions.

The level of knowledge among employers is not satisfactory when it comes to the risks of discrimination associated with the use of AI and other types of automated decision-making. Another conclusion is therefore that employers must take greater responsibility for combating

discrimination when using AI and other types of automated decision-making by complying with the Discrimination Act. Employers are obliged to comply with the requirement under the Discrimination Act to take active measures, regardless of which technical solutions they are using.

In conclusion, we are able to establish that the development of AI and other automated decision-making is an area that DO and other European equality bodies need to monitor in order to contribute to the prevention of discrimination.

Kapitel 1 Inledning

Användandet av artificiell intelligens (AI) och annat automatiserat beslutsfattande i arbetslivet har ökat kraftigt under de senaste decennierna. De främsta drivkrafterna bakom denna utveckling är ökad kostnadseffektivitet och produktivitet i relation till utförandet av specifika arbetsuppgifter eller utmaningar som företag och organisationer står inför. Inte sällan framhålls också att AI och annat automatiserat beslutsfattande är ett sätt att uppnå bättre kvalitet och ökad rättvisa i organisatoriskt beslutsfattande.[11] Systemen ska öka graden av objektivitet, konsekvens och rättvisa i arbetslivsprocesser vid exempelvis rekryteringsbeslut, prestationsutvärderingar och befordringar.[12]

Sveriges nationella inriktning för AI beskriver att Sverige ska vara ledande i att använda AI och att detta behöver göras på ett etiskt, säkert och hållbart sätt.[13] Samtidigt har flera europeiska likabehandlingsorgan uppmärksammat att tillämningen av AI och annat automatiserat beslutsfattande kan innebära risker för diskriminering och utgöra hinder för lika rättigheter och möjligheter. Det europeiska nätverket för likabehandlingsorgan, Equinet, har lyft fram att bland annat bristande insyn (transparens) och svarta lådan-fenomenet, det vill säga datorprogram som självständigt gör kopplingar och ser mönster och samband mellan olika data, innebär risker för diskriminering.[14] Equinet betonar att icke-diskriminering och lika rättigheter och möjligheter bör bli en naturlig del i all utveckling av AI och annat automatiserat beslutsfattande och att likabehandlingsorgan behöver få

11 Shestakova (2021).
12 Drage och Mackereth (2022).
13 Regeringen (2018)
14 Equinet (2020a; 2020b; 2020c).

tillgång till de resurser och möjligheter som krävs för att ha en effektiv tillsyn inom området.[15]

Equinet belyser även vilka förödande konsekvenser som felaktig användning av AI-system fick i Nederländerna.[16] Där använde sig Skattemyndigheten felaktigt av självlärande algoritmer vilket resulterade i diskriminering. Efter att ha infört skärpt lagstiftning för att bekämpa bedrägerier gällande barnbidrag, använde Skattemyndigheten ”utländskt medborgarskap” som en (negativ) riskfaktor. Dessutom registrerade Skattemyndigheten olagligt om personer som inte var nederländska medborgare också hade ett annat medborgarskap, och använde sig av dessa nationalitetsuppgifter i sina bedrägeriutredningar. Det ledde till att föräldrar som omfattades av dessa utredningar i vissa fall tvingades betala tillbaka stora bidrag. En parlamentarisk utredning kom 2020 fram till att det sätt på vilket föräldrarna behandlades utgjorde en ”aldrig tidigare skådad orättvisa”, vilket ledde till att regeringen avgick.

Europeiska unionens byrå för grundläggande rättigheter (FRA), har uppmärksammat att användare av AI och annat automatiserat beslutsfattande ofta är omedvetna om de risker för diskriminering som kan uppstå.[17] Även Diskrimineringsombudsmannen (DO) har i en tidigare rapport konstaterat att kunskapsnivån om risker för diskriminering och hinder för lika rättigheter och möjligheter inte är tillfredställande bland de myndigheter som använder tekniken för att fatta beslut gentemot enskilda.[18]

15 Equinet (2020a; 2020b; 2020c).
16 Equinet (2022).
17 Se för exempel Europeiska unionens byrå för grundläggande rättigheter (2020).
18 Diskrimineringsombudsmannen (2022).

Regeringen har gett DO i uppdrag att fördjupa kunskapen om risker för diskriminering vid användandet av AI och annat automatiserat beslutsfattande inom arbetslivet.[19]

Syftet med uppdraget

Enligt regeringsuppdraget ska DO kartlägga vilka risker för diskriminering inom arbetslivet som användandet av AI och annat automatiserat beslutsfattande kan medföra och i vilken utsträckning och i vilka sammanhang som arbetsgivare använder sådana tekniska lösningar.[20] Vidare anges i uppdraget att det bör belysa kunskapsnivån hos arbetsgivare om förhållanden som kan leda till diskriminering när det gäller AI och annat automatiserat beslutsfattande. Slutredovisningen av regeringens uppdrag till DO består av denna rapport. I arbetet med uppdraget har DO arbetat med följande:

1. Kartlägga vilka risker för diskriminering som användandet av AI och annat automatiserat beslutsfattande kan medföra i arbetslivet.
2. Undersöka i vilken utsträckning och i vilka sammanhang som arbetsgivare använder sådana tekniska lösningar.
3. Belysa kunskapsnivån hos arbetsgivare om förhållanden som kan leda till diskriminering när det gäller AI och annat automatiserat beslutsfattande.

Vi har i arbetet med uppdraget i första hand fokuserat på hur AI och annat automatiserat beslutsfattande används i samband med rekrytering och i arbetsledningen av personal.

19 Arbetsmarknadsdepartementet (2022), Regeringsbeslut 2022-06-07, A2022/00877.
20 Arbetsmarknadsdepartementet (2022), Regeringsbeslut 2022-06-07, A2022/00877.

Disposition

Rapporten ska läsas som en antologi och är indelad i olika kapitel.

I kapitel 1 ger vi en bakgrund till uppdraget, syfte och metod. Vi redogör även för definitioner av begreppet AI baserat på underlag från forskarna Stefan Larsson och James White, Institutionen för teknik och samhälle, båda vid Lunds Tekniska Högskola.

I kapitel 2 beskriver vi regleringen av AI på europeisk nivå.

Kapitel 3 ger en överblick över diskrimineringslagen

Kapitel 4 består av en forskningsöversikt. Kapitlet redogör för det aktuella forskningsläget gällande risker för diskriminering vid tillämpningen av AI i arbetslivet, möjligheter att förebygga diskriminering i arbetslivet och vilka kunskapsluckor som finns inom forskningsfältet.

Kapitel 5 redovisar resultat och analys av en telefonenkät riktad till rekryteringsföretag och större företag i privat sektor med fler än 100 anställda. Kapitlet beskriver i vilken utsträckning och i vilka sammanhang som arbetsgivare i Sverige använder sig av AI och annat automatiserat beslutsfattande i arbetslivet. I kapitlet resonerar också forskarna kring hur medvetna arbetsgivare är om vilka risker för diskriminering som kan finnas och vad de gör för att minimera riskerna.

I det sjätte och avslutande kapitlet, sammanfattar DO resultaten och presenterar övergripande slutsatser om risker för diskriminering vid tillämpning av AI och annat automatiserat beslutsfattande i arbetslivet.

Metod

Rapporten baseras i huvudsak på två kunskapsunderlag som har skrivits av forskare på uppdrag av DO och på ett antal bakgrundsintervjuer som DO har genomfört med arbetsgivarorganisationer. DO vill förtydliga att de forskargrupper som DO har anlitat har i delar använt en annan definition av diskriminering än diskrimineringslagens definition.

Forskningsöversikt över forskning om risker för diskriminering vid tillämpning av AI och annat automatiserat beslutsfattande i arbetslivet

DO har anlitat forskarna Martin Wolgast, biträdande prefekt, och Sima Nourali Wolgast, universitetslektor, båda vid Institutionen för psykologi på Lunds universitet, för att ta fram en systematisk forskningsöversikt. Forskningsöversikten baseras på 58 forskningsartiklar och inkluderar även internationell forskning.[21] Den fokuserar dels på det aktuella forskningsläget gällande risker för diskriminering vid användning av AI och annat automatiserat beslutsfattande i arbetslivet, dels på användning av AI och annat automatiserat beslutsfattande som kan förebygga diskriminering i arbetslivet samt vilka kunskapsluckor som finns inom forskningsfältet. Forskningsartiklarna har publicerats under perioden 2017–2022 och sökningarna har inte begränsats till ett visst geografiskt område. I bilaga 1 finns mer information om metoden.

Undersökning av hur arbetsgivare använder sig av AI och annat automatiserat beslutsfattande och vilken kunskap de har om risker för diskriminering

DO har anlitat Research Institute for Sustainable AI (RISAI) och forskarna Stefan Larsson, lektor och docent i teknik och social förändring, och James White, forskare vid institutionen för teknik och samhälle, Lunds Tekniska Högskola, Lunds universitet. Tillsammans med kunskaps- och analysföretaget Novus har de genomfört telefonenkäter

21 Se metodbilaga 1.

med 10 rekryteringsföretag och 100 företag med fler än 100 anställda. Syftet har varit att undersöka vilka risker för diskriminering som finns inbyggda i tekniska lösningar baserade på AI och annat automatiserat beslutsfattande i arbetslivet och vilken kunskap arbetsgivare har om dessa risker och vad de gör för att minimera dem. Stefan Larsson och James White har även skrivit grunden till avsnittet AI – ett svårfångat begrepp i rapportens första kapitel som DO har reviderat. I bilaga 2 finns mer information om metoden.

Stefan Larsson har i egenskap av forskare inom området även lämnat synpunkter på DO:s slutgiltiga manus.

Bakgrundsintervjuer med arbetsgivarorganisationer

För att få en generell bild av kunskapsnivån hos arbetsgivare om AI och annat automatiserat beslutsfattande har DO genomfört fyra bakgrundsintervjuer med centrala arbetsgivarrepresentanter från vitt skilda grenar i arbetslivet: Arbetsgivarverket, Bankinstitutens Arbetsgivarorganisation, Sveriges Kommuner och Regioner samt Tech Sverige.

AI – ett svårfångat begrepp

Det är viktigt att beskriva definitionen av begreppet AI och annat automatiserat beslutsfattande och hur den ser ut i olika kontexter och förändras över tid för att kunna förstå problematiken med eventuella risker för diskriminering kopplat till AI och annat automatiserat beslutsfattande i arbetslivet.

I den här rapporten har de forskare som DO har anlitat valt att använda något olika definitioner av AI och annat automatiserat beslutsfattande. Det i sig har inga större konsekvenser för de slutsatser som kan dras. I stället speglar det svårigheterna med att kartlägga ett område där själva definitionen är undanglidande och ständigt i rörelse i takt med teknisk utveckling och tillämpning.

I kapitel 4 där forskningsöversikten presenteras blir det tydligt att definitionerna av AI och annat automatiserat beslutsfattande skiljer sig åt mellan olika forskningsartiklar. I det avseendet är det svårt att veta om forskningsartiklarna ens studerar samma sak. I kapitel 5, som undersöker användningen av AI och annat automatiserat beslutsfattande i arbetslivet, har forskarna som genomfört studien antagit ett pragmatiskt förhållningssätt till definitionen av AI och annat automatiserat beslutsfattande. Det innebär att forskarna inte har gett en strikt definition utan att de i stället har beskrivit för respondenterna vad AI kan innebära. Det har gjort det möjligt att tillåta att respondenternas skiftande förståelse för vad AI och annat automatiserat beslutfattande är inkluderas i undersökningen.

Definitionen av AI har visat sig svårfångad såväl som forskningsfält som för den europeiska lagstiftaren.[22] I mitten av 2018 utnämnde EU-kommissionen 52 medlemmar till en AI-expertgrupp. Gruppen fick i uppdrag att ta fram råd om AI-risker och etik, som senare skulle bidra till utvecklingen av EU:s föreslagna regelverk för AI. Som en del av detta arbete lade AI-expertgruppen fram följande definition på AI:

> Artificiella intelligenssystem (AI) är programvarusystem (och eventuellt även hårdvarusystem) som har konstruerats av människor och som när de får ett komplext mål agerar i den fysiska eller digitala dimensionen genom att uppfatta sin omgivning via datainsamling och att tolka insamlade strukturerade eller ostrukturerade data, resonerar om den kunskap eller behandlar den information som härletts ur denna data och beslutar om den bästa åtgärd eller de bästa åtgärderna som ska vidtas för att uppnå det fastställda målet. AI-system kan använda symboliska regler eller lära sig en numerisk modell. De kan också anpassa sitt beteende genom att analysera hur miljön har påverkats av deras föregående åtgärder.[23]

22 Larsson och Ledendal (2022).
23 AI HLEG (2019b) En definition av AI.

Denna definition innehåller två aspekter som är viktiga att utveckla, inte minst när vi talar om risker för diskriminering i samband med användandet av AI.

För det första omfattar definitionen både strukturerade och ostrukturerade data, det vill säga både AI-metoder där informationen som används är organiserad i en förutbestämd struktur, och metoder där informationen inte är organiserad på något förutbestämt sätt. Detta pekar på en viktig skillnad i maskininlärning mellan metoder som identifierar mönster inom ostrukturerade data (det vill säga så kallad oövervakad inlärning) och metoder som producerar statistiska modeller baserade på strukturerade och märkta data, så kallad övervakad inlärning. Utan behov av förberedande arbete för att rensa och organisera data kan oövervakad inlärning utföras på väldigt stora dataset för att skapa kopplingar och producera modeller som kan uttrycka dessa kopplingar på nya sätt.

Den publika chattbotten ChatGPT är ett exempel på stora språkmodeller som är tränade på stora mängder internettext. Den kan producera texter av olika slag i mångmiljonskala som upplevs som trovärdiga av mänskliga läsare. Övervakad inlärning utförs ofta på mindre och mer noggrant kontrollerade databaser för att göra förutsägelser om liknande men tidigare osedd data. Det finns till exempel modeller som tränas på taggade bilder av människor som medvetet gör ansiktsuttryck med målet att modellen ska kunna känna igen dessa uttryck i andra bilder. Dessa två metoder har olika funktioner och är förknippade med olika risker.

För det andra inkluderar definitionen ett tekniskt inriktat fokus på data, programvara och hårdvara. Även om det är viktigt att inte begränsa AI-system till sin kod är det också viktigt att inse vad denna definition utesluter. Alla AI-modeller måste produceras innan de kan tillämpas. Det innebär förberedelse av data, val av lämpliga metoder och upprepad justering av parametrar för att förbättra modellens kapacitet. Det arbetet utförs i första hand av människor. De data som behövs för att träna och

tillämpa en AI-modell är resultatet av beslut om vad som är viktigt för AI att veta och kan utgöras av en representation av inneboende sociala fördomar och ojämlikheter.

AI används för att referera till en mängd olika datoralgoritmer – det vill säga sekvenser av regler eller instruktioner – som har programmerats att agera på insamlad data på komplicerade men förutsägbara sätt. Dessa algoritmer kan fatta beslut mycket mer effektivt än människor och är outtröttliga. De kan kombineras med andra algoritmer, inkluderas i automatiserade beslutssystem och bäddas in i maskiner som kan röra sig och vidta åtgärder i världen – många autonoma system, inklusive robotgräsklippare och självkörande bilar, använder sig av AI-system. Även om de i princip är förutsägbara, kan dessa större och mer komplexa AI-system i praktiken fatta oväntade beslut, innehålla fel och vidta oönskade åtgärder, vars yttersta orsak kan vara svår att fastställa även för dem med den specialiserade kunskap som krävs för att göra det.

På senare år har de AI-metoder som ryms inom maskininlärning kommit att få en stor betydelse. Det finns många metoder för maskinlärning. En av de mest omtalade under det senaste året handlar om så kallade transformermodeller som använder stora mängder data. Den nuvarande trenden inom området går mot mer data och mer beräkning, vilket oundvikligen minskar förståelsen för hur systemet fungerar. Flerskiktade neurala nätverk, exempelvis, där mycket av innovationen under de senaste tio åren har gjorts, har ytterligare ett problem genom att de inte bara är komplexa utan också oförmögna att rapportera tillbaka till utvecklaren om hur de når ett visst beslut. Detta är känt som ett förklarbarhetsproblem, vilket komplicerar frågor om ansvarsskyldighet när saker går fel.[24]

24 Se Larsson och Heintz (2020).

Medan de som arbetar inom branschen ofta är försiktiga med hur de hänvisar till AI-algoritmer, modeller och system, och kan förväntas ha tagit hänsyn till var gränserna för deras kunskap och förståelse ligger, är detta inte nödvändigtvis fallet hos allmänheten i stort. Här används terminologierna på olika, inkonsekventa och ospecifika sätt, utan åtskillnad mellan underliggande tillvägagångssätt och deras inneboende styrkor och svagheter. Detta kompliceras ytterligare av medvetna missbruk och överdrifter eller strategiska sätt att använda terminologin för att dra nytta av förhoppningar och förväntningar för sektorn.[25]

Algoritmen är inte en steg-för-steg-process mot ett förutbestämt resultat, utan beskriver snarare hur ett system kan lära sig att nå sina mål på egen hand. Maskininlärning täcker ett brett spektrum av metoder, inklusive men inte begränsat till djupinlärning, neurala nätverk och så kallat generativ AI, som tillämpas på ett brett spektrum av problem eller områden. Djupinlärning är en klass av algoritmer för maskininlärning som använder flera lager för att successivt extrahera funktioner på högre nivå från underliggande data.[26] Vid bildbehandling kan till exempel lägre lager identifiera kanter, medan högre lager kan identifiera begrepp som är relevanta för en människa, såsom siffror eller bokstäver eller ansikten. Moderna varianter är ofta komplexa så kallade neurala nätverk.[27] Det finns många metoder för detta maskinlärande.

DO har i tidigare kartläggning av statliga myndigheters kunskap om risker för diskriminering vid tillämpning av AI[28] använt en definition som speglar EU-kommissionens formuleringar i det som betraktas som

25 Se till exempel artiklar i Medium (20 juni 2022) och Medium (18 april 2022). För en mer samhällsvetenskaplig analys av hur branschen förstår och hanterar hype se Slota med flera (2020).
26 Se Russell och Norvig (2022).
27 Larsson och Heintz (2020).
28 Diskrimineringsombudsmannen (2022).

EU:s AI-strategi.[29] Den definitionen användes även initialt av dess AI-expertgrupp[30], och förekommer i rapporter från Myndigheten för digital förvaltning (DIGG) inom det nationella AI-uppdraget.[31] En utmaning är sammantaget att AI-begreppet skiljer sig åt – vilket har påpekats i studier – om man jämför datavetenskap, juridik och allmänhetens uppfattning.[32]

Det är även svårt att fastställa definitionen av automatiserat beslutsfattande. I studier riktade mot offentlig förvaltning används ofta det engelska uttrycket automated-decision making, med förkortningen ADM. Det är ett paraplybegrepp som också inrymmer delvis automation som beslutsstöd och olika grader av mänsklig kontroll.[33] Även enklare men algoritmiskt regelstyrda system för beslutsfattande kan därmed inkluderas, vilket också gör att det finns en längre historik kring användningen av automatiserade beslut.[34] Skillnaden mellan beslutsstöd och beslut kan dock vara avgörande, inte minst eftersom undersökningar av vad människor litar på för typ av automation hos offentlig sektor visar att förtroendet är långt mycket högre för automatiserade beslutsstöd än helt automatiserat beslutsfattande.[35] Sammantaget kan vi konstatera att det finns skilda definitioner av AI och annat automatiserat beslutsfattande och att det i sin tur gör det svårt att studera förekomsten av teknikerna. För analysen om risker för diskriminering är dock den exakta definitionen av mindre betydelse.

29 Europeiska kommissionen (2018).
30 AI HLEG (2019c).
31 Myndigheten för digital förvaltning, DIGG (2020).
32 Se Larsson och Heintz (2020).
33 Roehl (2022).
34 Riksrevisionen (2020) konstaterar till exempel att automatiserade beslut är vanliga i statsförvaltningen och har funnits sedan 1970-talet.
35 Insight Intelligence (2022).

Kapitel 2 Reglering med sikte på AI på europeisk nivå

Det finns anledning att kort nämna två europeiska initiativ till reglering av AI, som också berör diskrimineringsfrågor, även om detta uppdrag utgår från svensk diskrimineringslagstiftning. Initiativen kommer av allt att döma påverka eller få betydelse för AI-system och digitala tjänster som erbjuds av väldigt stora teknikaktörer. Båda initiativen inkluderar bestämmelser som tar sikte på rätten till icke-diskriminering.

I EU pågår arbetet med att ta fram en förordning om harmoniserade regler för AI, ofta kallad AI-förordningen, som ska utgöra en rättslig ram för tillförlitlig AI inom EU.[36]

EU-kommissionen lämnade i april 2021 ett förslag till AI-förordning, och i slutet av år 2022 och början av år 2023 lämnade Europeiska unionens råd och EU-parlamentet sina förslag till ändringar i EU-kommissionens förslag till AI-förordning. Under sommaren och hösten 2023 pågår i skrivande stund förhandlingar mellan EU-kommissionen, Europeiska unionens råd och EU-parlamentet för att enas om den slutliga utformningen av förordningen. AI-förordningen kommer att bli rättsligt bindande för och direkt tillämplig i alla EU:s medlemsländer.

36 Se exempelvis Europaparlamentet Nyheter (2023).; Europakommissionen (2023).

Förslaget till AI-förordning innehåller vissa bestämmelser som tar sikte på skyddet av de så kallade grundläggande rättigheterna i EU, och en av dessa är rätten att inte bli diskriminerad (artikel 21 i Europeiska unionens stadga om de grundläggande rättigheterna).[37] I förslaget till AI-förordning finns bland annat förslag om att vissa AI-system ska godkännas innan de får släppas ut på den europeiska marknaden och att det i vissa fall kan krävas att aktörer som använder AI-system registrerar det i en offentlig databas. Ett förslag är också att tillsynsmyndigheter inrättas, och att både dessa tillsynsmyndigheter och organ som övervakar de grundläggande rättigheterna ska få tillgång till information och dokumentation om AI-system för att kunna utöva tillsyn respektive kunna bevaka rätten till bland annat icke-diskriminering.[38]

Equinet har ställt sig positivt till flera av de förslag till förändringar som Europeiska rådet och EU-parlamentet har föreslagit, men har också understrukit vikten av att likabehandlingsorgan får tillgång till begriplig information och att det behövs samarbete mellan likabehandlingsorgan och de tillsynsmyndigheter som ska övervaka efterlevnaden av AI-förordningen.[39]

Det andra europeiska initiativet till reglering av AI är Europarådet som i skrivande stund arbetar med att ta fram en konvention om AI och mänskliga rättigheter.[40]

37 Europeiska unionens officiella tidning (2012).
38 Europaparlamentet (2023).
39 Equinet (2023)
40 Europarådet (2023).

Kapitel 3 Diskrimineringslagen

Det är viktigt att ta avstamp i vad som avses med diskriminering enligt diskrimineringslagen (2008:567) för att kunna bedöma risker för diskriminering vid användningen av AI och annat automatiserat beslutsfattande i arbetslivet. Diskrimineringslagen är tillämplig vid användningen av AI och automatiserat beslutsfattande, vilket DO har konstaterat i tidigare undersökningar.[41]

Diskrimineringslagen syftar till att motverka diskriminering och på andra sätt främja lika rättigheter och möjligheter oavsett kön, könsöverskridande identitet eller uttryck, etnisk tillhörighet, religion eller annan trosuppfattning, funktionsnedsättning, sexuell läggning eller ålder.[42]

I rapporten fokuserar vi på risker för diskriminering inom arbetslivet. Enligt diskrimineringslagen får en arbetsgivare inte diskriminera den som hos arbetsgivaren är arbetstagare, gör en förfrågan om eller söker arbete, söker eller fullgör praktik eller står till förfogande för att utföra eller utför arbete som inhyrd eller inlånad arbetskraft.[43]

Med diskriminering menas i diskrimineringslagen, lite förenklat, att en person missgynnas eller kränks och att det finns ett samband mellan detta missgynnande och någon av de sju diskrimineringsgrunderna. Vidare måste missgynnandet eller kränkningen ha skett inom något av de samhällsområden som lagen omfattar, exempelvis i arbetslivet, i utbildningssystemet, inom hälso- och sjukvården eller vid tillhandahållande av bostäder till allmänheten.

41 Se Diskrimineringsombudsmannen 2022.
42 1 kapitlet 1 § diskrimineringslagen.
43 2 kapitlet 1 § diskrimineringslagen.

Andra kapitlet i diskrimineringslagen behandlar diskrimineringsförbuden inom olika områden. Det finns sex former av diskriminering: direkt diskriminering, indirekt diskriminering, trakasserier, sexuella trakasserier, bristande tillgänglighet och instruktioner att diskriminera.[44] Tredje kapitlet i diskrimineringslagen handlar om aktiva åtgärder. Det innebär att arbetsgivare ska bedriva ett främjande och förebyggande arbete för att motverka diskriminering och på annat sätt verka för lika rättigheter och möjligheter.

Direkt diskriminering: diskriminerande behandling

Direkt diskriminering är när någon missgynnas genom att behandlas sämre än någon annan i en jämförbar situation. Ett missgynnande är om någons handlande, beslut eller beteende får en negativ följd för någon annan, det vill säga en skada, nackdel eller ett obehag. Avgörande är att en negativ effekt inträder, inte vilken orsak som kan ligga bakom missgynnandet.[45] För att ett sådant missgynnande ska utgöra diskriminering måste också en jämförelse göras med hur någon annan person har behandlats eller skulle ha behandlats i en jämförbar situation. Slutligen måste det finnas ett orsakssamband mellan den diskriminerande handlingen och en av diskrimineringsgrunderna.[46]

För att en händelse ska utgöra direkt diskriminering behöver alltså tre kriterier vara uppfyllda. Det första kriteriet är att någon ska ha missgynnats, det vill säga behandlats sämre eller utsatts för någon form av förlust eller nackdel. Det andra kriteriet jämförbar situation innebär att detta missgynnande måste jämföras med hur andra behandlas, har

44 1 kapitlet 4 § diskrimineringslagen. Diskrimineringsformerna trakasserier, sexuella trakasserier, bristande tillgänglighet och instruktioner att diskriminera tas inte upp vidare i denna rapport.
45 Proposition 2007/08:95, sida 486–487.
46 1 kapitel 4 § 1 diskrimineringslagen.

behandlats eller skulle ha behandlats i en jämförbar situation. Det kan exempelvis handla om att undersöka hur andra personer med likvärdiga meriter och kvalifikationer som har sökt ett visst arbete har behandlats i en rekryterings- och urvalssituation. Om det saknas en verklig person att jämföra med kan jämförelsen göras med hur en hypotetisk, förmodad, person skulle ha blivit behandlad. Det tredje och sista kriteriet för direkt diskriminering är att det måste finnas ett orsakssamband mellan missgynnandet och någon av diskrimineringsgrunderna. Det räcker att diskrimineringsgrunden är en av flera orsaker till att någon blir missgynnad. Sambandet kan vara starkt eller svagt. Allra starkast är det om det finns en avsikt att missgynna en person på grund av någon av diskrimineringsgrunderna. En diskriminerande avsikt är emellertid inte nödvändig för att diskriminering ska kunna vara för handen.[47]

Indirekt diskriminering: diskriminerande kriterier eller förfaringssätt

Indirekt diskriminering innebär att någon missgynnas genom tillämpning av en bestämmelse, ett kriterium eller ett förfaringssätt som framstår som neutralt men som kan komma att särskilt missgynna personer med visst kön, viss könsöverskridande identitet eller uttryck, viss etnisk tillhörighet, viss religion eller annan trosuppfattning, viss funktionsnedsättning, viss sexuell läggning eller viss ålder. Detta såvida inte bestämmelsen, kriteriet eller förfaringssättet har ett berättigat syfte och de medel som används är lämpliga och nödvändiga för att uppnå syftet.[48] Ett exempel på indirekt diskriminering skulle kunna vara att en arbetsgivare kräver mycket goda kunskaper i svenska för en tjänst där språkkunskaper inte är avgörande för möjligheterna att utföra det arbete som tjänsten innefattar.[49]

47 Proposition 2007/08:95 sida 486–490.
48 4 kapitel 1 § 2 diskrimineringslagen.
49 Proposition 2007/08:95 sida 492.

Aktiva åtgärder för att motverka diskriminering

Aktiva åtgärder är ett förebyggande och främjande arbete för att inom en verksamhet motverka diskriminering och på annat sätt verka för lika rättigheter och möjligheter oavsett kön, könsöverskridande identitet eller uttryck, etnisk tillhörighet, religion eller annan trosuppfattning, funktionsnedsättning, sexuell läggning eller ålder.[50] Arbetsgivare ska enligt diskrimineringslagen genomföra aktiva åtgärder för att motverka diskriminering inom fem områden: arbetsförhållanden, bestämmelser och praxis om löner och andra anställningsvillkor, rekrytering och befordran, utbildning och övrig kompetensutveckling, och möjligheter att förena förvärvsarbete med föräldraskap. Arbetet med aktiva åtgärder ska genomföras fortlöpande i följande fyra steg:

1. undersöka om det finns risker för diskriminering eller repressalier eller om det finns andra hinder för enskildas lika rättigheter och möjligheter i verksamheten
2. analysera orsaker till upptäckta risker och hinder
3. vidta de förebyggande och främjande åtgärder som skäligen kan krävas
4. följa upp och utvärdera arbetet.[51]

50 3 kapitel 1 § diskrimineringslagen.
51 3 kapitel 2 § diskrimineringslagen.

Kapitel 4 Forskningsöversikt: risker för diskriminering i arbetslivet

Detta kapitel redogör för resultaten i den forskningsöversikt som Martin Wolgast och Sima Nourali Wolgast vid Institutionen för psykologi vid Lunds universitet har sammanställt på uppdrag av DO. Kapitlet bygger på ett manus som DO har redigerat.

Forskningsöversikten ger en nulägesbild av forskningsläget om risker för diskriminering i arbetslivet utifrån de avgränsningar som har gjorts i forskarnas sökningar efter artiklar. Forskning utvecklas ständigt och nya resultat publiceras kontinuerligt och bidrar till kunskapsutvecklingen inom området. Forskning om diskriminering och AI och annat automatiserat beslutsfattande sker i många olika vetenskapliga discipliner och publiceras i en rad olika tidskrifter med en stor variation gällande terminologi och begreppsanvändning. Som läsare är det därför viktigt att vara medveten om att det kan finnas exempel på relevant forskning som inte har fångats upp i den aktuella forskningsöversikten.

Syfte, frågeställningar och metod

Forskningsöversikten syftar till att besvara följande frågor:

1. Hur ser det aktuella forskningsläget ut gällande riskerna för diskriminering vid användning av AI och annat automatiserat beslutsfattande i arbetslivet?
2. Hur ser det aktuella forskningsläget ut gällande användning av AI och annat automatiserat beslutsfattande som kan förebygga diskriminering i arbetslivet?
3. Vilka är kunskapsluckorna inom forskningsfältet?

Forskningsöversikten baseras på resultat publicerade i vetenskapliga tidskrifter med kollegial granskning (så kallad peer-review) under perioden 2017–2022. Forskningsöversikten omfattar endast forskning om sådana system som bygger på AI eller andra former av automatiserat beslutsfattande. Nästan alla artiklar som har tagits med i forskningsöversikten handlar dock om beslutsfattande baserat på maskininlärningsalgoritmer (machine learning) – vissa mer styrda och övervakade och andra mer öppna. Forskning som rör exempelvis datoriserade tester eller spel som urvalsinstrument eller användandet av sociala medier för att nå ut till och bedöma potentiella kandidater till tjänster har alltså enbart inkluderats om de undersökta systemen bygger på AI eller andra former av automatiserat beslutsfattande.

På liknande sätt har bara studier som rör möjligheter vid användning av AI-system och risker för diskriminering i arbetslivet i samband med användandet av AI och annat automatiserat beslutsfattande inkluderats. Detta innebär att det som har exkluderats är forskning rörande andra effekter i arbetslivet till följd av den ökade användningen av system baserade på AI och annat automatiserat beslutsfattande. Det gäller till exempel forskning om att vissa typer av arbeten kan försvinna eller att farliga arbetsuppgifter kan komma att utföras av maskiner framöver, om inte denna forskning också på ett tydligt sätt kan kopplas till möjligheter vid användning av AI-system och risker för diskriminering.

Detta innebär att översikten inte inkluderar forskning som rör sådana former av digitalisering eller datorunderstödda processer som inte bygger på AI och annat automatiserat beslutsfattande. Detta innefattar exempelvis att genomföra olika former av tester online i en urvalssituation, användandet av digitala anställningsintervjuer eller utvecklingen av nya urvalsmetoder, till exempel spel som mäter problemlösningsförmåga, simultankapacitet eller andra färdigheter.[52]

52 Woods, Ahmed, Nikolaou, Costa och Anderson (2020).

De artiklar som har inkluderats i forskningsöversikten utgår inte från de lagstadgade diskrimineringsgrunderna, eftersom den vetenskapliga litteraturen inte alltid använder samma begrepp och klassifikationer som i diskrimineringslagen.

Sammanlagt ingår 58 artiklar i det slutgiltiga underlaget som ligger till grund för resultatsammanställningen i denna rapport. Dessa artiklar kommer från olika discipliner och vetenskapsområden, alltifrån sociologi och juridik till datavetenskap. De flesta artiklar diskuterar olika former av risker för diskriminering i samband med att AI och annat automatiserat beslutsfattande tillämpas i arbetslivet, men det finns också många studier som lyfter fram möjligheter vid användning av AI-system och fördelar eller som pekar på både risker och möjligheter. Vidare är en relativt stor andel (cirka 55 procent) av artiklarna antingen av konceptuell natur, det vill säga de för principiella eller teoretiska resonemang om risker eller möjligheter med system baserade på AI och annat automatiserat beslutsfattande, eller översiktsartiklar som sammanfattar och diskuterar forskning på området som har genomförts av andra. För mer information om metod och tillvägagångsätt se bilaga 1.

AI och annat automatiserat beslutsfattande och dess tillämpningsområden

Kapitlet inleds med en grundläggande beskrivning av vad AI och annat automatiserat beslutsfattande är och vilka tillämpningsområden dessa system har fått i ett arbetslivssammanhang. Detta för att underlätta förståelsen av de resonemang och resultat som förekommer i forskningsöversikten.

AI-processer: maskiner som används för att ersätta eller förbättra mänskligt beslutsfattande

Begreppen AI och automatiserat beslutsfattande är inte helt lätta att definiera, vilket har konstaterats i tidigare kapitel. De används dock ofta som samlingsbegrepp för en rad olika fenomen relaterade till när digital teknologi används för att skapa system som kan utföra arbetsuppgifter som man tidigare har antagit kräver mänsklig intelligens.[53] AI och automatiserat beslutsfattande kan alltså betraktas som övergripande sätt att benämna processer där maskiner används för att ersätta eller förbättra mänskligt beslutsfattande.[54]

Det åsyftas ofta någon form av procedur som bygger på den underkategori av AI som brukar kallas maskininlärning när AI och annat automatiserat beslutsfattande diskuteras och tillämpas i relation till arbetslivet. Med maskininlärning avses datoriserade system som på ett självständigt sätt förbättrar sin förmåga att över tid utföra en specifik uppgift genom att systemet lär sig av sina ackumulerade erfarenheter.[55] Lite förenklat kan man alltså säga att det rör sig om metoder för att med befintlig data träna datoriserade system att upptäcka och lära sig regler för att lösa en uppgift, utan att systemen har programmerats med regler som på ett mer exakt sätt preciserar hur de ska utföra just den uppgiften.

Det är viktigt att etablera en grundläggande förståelse för hur de system och metoder som diskuterats ovan fungerar för att förstå både möjligheter vid användning av AI-system och risker med AI och annat automatiserat beslutsfattande i relation till diskriminering i arbetslivet. På ett övergripande plan beskrivs ofta AI – och då i synnerhet maskininlärning – som system utformande för att göra prediktioner och klassifikationer.[56]

53 Office for AI (2019).
54 Shestakova (2021).
55 Mitchell (1997).
56 Agrawal, Gans och Goldfarb (2018); Birhane (2021).

Grunden i automatiserat beslutsfattande och beslutsfattande baserat på AI utgörs då av de så kallade algoritmer som systemen är uppbyggda kring. Med algoritmer avses en systematisk procedur som specificerar hur en beräkning ska genomföras eller ett visst problem ska lösas.[57] Algoritmbaserat beslutsfattande innebär alltså att beslutsfattandet bygger på någon form av beräkningsmodell som fattar beslut baserat på regler och statistiska metoder utan direkt mänsklig inblandning eller styrning.[58] Systemen – eller ”modellerna” som de brukar kallas i dessa sammanhang – behöver skapas och etableras genom en process som består av flera olika steg för att kunna generera dessa beslut eller prediktioner.

Initialt måste den data som modellen ska lära sig utifrån framställas och förberedas. Denna data är ofta omfattande och består av en stor mängd information om en rad olika förhållanden och faktorer. I arbetslivssammanhang kan detta exempelvis handla om data i relation till en stor mängd tidigare rekryteringsbeslut, uppgifter om de anställda (rörande exempelvis utbildning, tidigare meriter, demografi och så vidare) inom en organisation, bransch eller på ett visst arbetsmarknadsområde, eller information om olika aspekter av prestation hos anställda. I detta steg är det alltså avgörande vilken data som samlas in, hur viktiga egenskaper eller variabler mäts och vilka grupper eller populationer som finns representerade i datamaterialet. I nästa steg ”tränas” modellen i relation till denna data genom att den analyserar och tar fram en optimal kombination av variabler som genererar så bra prediktioner som möjligt i relation till ett specifikt utfall, till exempel vilka egenskaper hos anställda som verkar förutsäga goda prestationer, att man stannar länge på tjänsten eller andra utfall av intresse.[59] Det är här av avgörande betydelse för modellens funktionssätt hur dessa utfall specificeras, vilka variabler som

57 Nationalencyklopedin, algoritm.
58 Lee (2018).
59 Larose och Larose (2014).

inkluderas i modellen och hur dessa variabler kombineras. Modellen testas sedan i relation till ny data för att utvärdera om den fungerar även i relation till ett underlag som är skilt från det den tränats på, varefter den implementeras och utvärderas på liknande sätt i praktisk tillämpning.[60]

AI och annat automatiserat beslutsfattande – ett sätt att minska kostnader, öka produktiviteten och öka kvalitet

De främsta drivkrafterna för den ökade användningen av AI och annat automatiserat beslutsfattande i arbetslivet har varit att reducera kostnader och tidsåtgång, öka produktiviteten, minimera olika former av risker samt att öka kvalitet och träffsäkerhet i viktiga beslut.[61]

Förutom dessa effektivitets- och produktivitetsskäl har AI och annat automatiserat beslutsfattande i arbetslivet även lyfts fram som ett sätt att åstadkomma minskat inflytande från mänskliga fördomar, föreställningar och partiskhet i organisatoriskt beslutsfattande. Tanken är här att systemen ska öka graden av objektivitet, konsekvens och rättvisa i centrala arbetslivsprocesser, exempelvis rekryteringsbeslut, prestationsutvärderingar och befordringar.[62]

De ovannämnda drivkrafterna har medfört att olika former av AI och annat automatiserat beslutsfattande i dag ses som nödvändiga för bevarad konkurrenskraft för många företag. Vidare har utbudet av olika former av produkter och tjänster som bygger på AI och annat automatiserat beslutsfattande i relation till centrala hr-frågor,

60 Kuhn och Johnson (2016).
61 Suen, Chen och Lu (2019); McDonald, Fisher och Connelly (2017); Woods, Ahmed, Nikolaou, Costa och Anderson (2020).
62 Drage och Mackereth (2022).

exempelvis rekrytering och prestationsmätningar, ökat kraftigt och fått omfattande spridning i arbetslivet.[63]

Den vetenskapliga litteraturen rörande AI och annat automatiserat beslutsfattande i arbetslivet har även den av förklarliga skäl varit föremål för en snabb ökningstakt under det senaste decenniet. Flera av dessa publikationer handlar om möjligheter vid användning av AI-system och risker med AI och annat automatiserat beslutsfattande i relation till diskriminering i arbetslivet och i relation till en rad olika centrala arbetsmarknadsprocesser. Merparten av studierna rör rekrytering och urval i vid bemärkelse, från hur tjänster riktas till specifika målgrupper eller hur så kallade passiva kandidater kan identifieras och kontaktas, till hur AI och annat automatiserat beslutsfattande kan användas vid anställningsintervjuer och slutligt urval. Därutöver finns det också flera studier som rör andra viktiga hr-områden, exempelvis prestations-uppföljning och beslut i relation till befordran eller lönesättning.

De flesta studier rör AI och annat automatiserat beslutsfattande som verktyg vid rekrytering och urval

Flertalet av de studier som har inkluderats i forskningsöversikten rör användandet av AI och annat automatiserat beslutsfattande i samband med rekrytering och urval inom arbetslivet. Rekrytering och urval ses i detta sammanhang som en process bestående av flera olika på varandra följande faser: från annonsering eller kommunikation av ett identifierat rekryteringsbehov till ett slutligt anställningsbeslut.[64] Inom detta fält har AI och annat automatiserat beslutsfattande använts på en rad olika sätt.

63 Daugherty och Wilson (2018); Köchling och Wehner (2020).
64 Chungyalpa och Karishma (2016).

Annonsering av lediga tjänster

Riktad annonsering för att nå rätt individer

Företag och organisationer använder sig alltmer av sociala medie-plattformar, exempelvis Facebook, Instagram och LinkedIn, för att annonsera tjänster och identifiera potentiella kandidater att anställa.[65] Dessa plattformar innefattar i sin tur algoritmbaserade modeller som exempelvis styr vilka användare som ser annonsen eller föreslår potentiella kandidater att kontakta för arbetsgivare. På detta sätt filtreras och riktas informationsflödet via AI och annat automatiserat beslutsfattande i syfte att föreslå lämpliga kandidater för organisationer som vill rekrytera, eller passande arbetsgivare att kontakta för personer som söker arbete. Grunden för dessa algoritmbaserade beslut utgörs av faktorer som exempelvis användarnas beskrivningar av sig själva, kriterier som användaren specificerat, deras tidigare val och beteenden hos andra liknande användare.[66]

Forskningsartiklarna lyfter fram flera fördelar med de nya sätten att nå ut med information om lediga tjänster eller etablera kontakt mellan arbetsgivare och potentiella anställda. Många av dessa fördelar har emellertid inte direkt med diskriminering att göra. Återkommande teman är exempelvis att systemen erbjuder kostnadseffektiva sätt att nå stora grupper, att tekniken möjliggör att arbetsgivare får tag i personer som de annars kanske inte skulle ha gjort, att tjänsten kan riktas till specifika målgrupper för att få så hög träffsäkerhet som möjligt samt att även så kallade ”passiva kandidater” – det vill säga personer som inte aktivt söker nytt arbete – kan nås. [67]

65 Roth, Bobko, Van Iddekinge och Thatcher (2016).
66 Simbeck (2019).
67 Horton (2017).

I relation till forskning om mer direkta möjligheter att förebygga diskriminering finns det en inkluderad studie som undersöker detta sätt att sprida annonser på. Studien är genomförd på den spanska arbetsmarknaden och undersöker annonsspridning med hjälp av verktyget InfoJobs. I denna studie fann forskarna inga skillnader i visningsfrekvens baserat på kön (man jämfört med kvinna) eller ålder (24 år jämfört med 38 år).[68]

Det finns även enstaka studier som pekar på att dessa metoder för att etablera kontakt mellan arbetsgivare och potentiella anställda inte medför undanträngningseffekter i relation till rekrytering av kandidater som inte har föreslagits av systemet.[69] Därutöver finns det principiella resonemang i publicerade studier som framför perspektivet att användandet av AI och annat automatiserat beslutsfattande inom detta område är ett sätt att nå ut till grupper och individer som inte nås i samma utsträckning via traditionella annonser eller som i vanliga fall inte ingår i organisationens kontaktsfär, men i de aktuella artiklarna undersöks inte detta empiriskt.[70]

Riktad annonsering exkluderar vissa arbetssökande

Forskning visar också att annonsering med stöd av AI och annat automatiserat beslutsfattande kan resultera i risker för diskriminering. Det kan leda till att potentiella arbetssökande exkluderas på ett sätt som har samband med någon av diskrimineringsgrunderna. Arbetsgivare kan exempelvis specificera att annonsen endast ska ses av användare i ett visst åldersintervall eller av de som har angett ett specifikt kön.[71] Det kan även ske om arbetsgivare anger visningskriterier som inte explicit rör någon av diskrimineringsgrunderna, till exempel om en annons ska visas för personer i vissa bostadsområden eller som bor inom ett visst avstånd från

68 Martínez, Vinas och Matute (2021).
69 Horton (2017).
70 Kim (2020); Horton (2017).
71 Kim och Scott (2018).

arbetsplatsen. Mot bakgrund av segregationen på bostadsmarknaden baserat på etnisk tillhörighet skulle ett sådant kriterium riskera att leda till diskriminering. [72] Algoritmerna styr visserligen inte bort från ett visst kön eller etnisk tillhörighet, men de styr bort från vissa geografiska områden. Det kan ändå innebära att vissa grupper riskerar att diskrimineras.

Risker kvarstår ändå för diskriminering beroende på funktionen hos de algoritmer som styr spridningen av annonserna, även om annonsen och de specifikationer som anges för dess spridning är både direkt och indirekt neutrala i relation till diskrimineringsgrunderna.[73] Under sådana omständigheter kan dessa algoritmer generera diskriminerande effekter genom de krav på kostnadseffektivitet som finns inbyggda i dem. Det innebär att algoritmerna försöker optimera visningen av annonserna på ett sätt som medför att annonsen ska spridas till så många intresserade användare som möjligt. En studie som rörde annonsering av tjänster inom naturvetenskap, teknik och matematik i 191 länder pekade på att annonserna i större utsträckning visades för män än för kvinnor, trots att annonsen och visningskriterierna skulle vara könsneutrala.[74] Denna skillnad uppkom på grund av det faktum att det var något dyrare att visa annonsen för kvinnor än för män, eftersom kvinnor generellt sätt tenderar att klicka på fler annonser än män, vilket gör dem till en något dyrare målgrupp.[75]

Algoritmerna riskerar att reproducera etablerade gruppskillnader

Skillnader i visningsfrekvenser för olika grupper kan uppkomma som ett resultat av de algoritmer för maskininlärning som finns inbyggda i plattformarnas system för AI och annat automatiserat beslutsfattande.

72 Kim och Scott (2018).
73 Goretzko och Finja Israel (2022).
74 Lambrecht och Tucker (2019).
75 Lambrecht och Tucker (2019).

Algoritmerna riskerar att agera på ett sätt som reproducerar etablerade gruppskillnader och mönster i denna data eftersom dessa algoritmer tränas i att fatta beslut med hjälp av historiska data, det vill säga hur plattformarnas användare har interagerat med (till exempel vad de har klickat på eller inte klickat på) liknande annonser tidigare.

Hur användarna hittills har interagerat med den aktuella annonsen utgör en del av dataunderlaget, vilket kan få drastiska effekter. I en experimentell studie genomförd i USA påvisades exempelvis märkbara skillnader baserade på hudfärg och kön, när algoritmerna strävade efter att visa annonserna på ett sätt som genererade så många klick som möjligt. Annonser inom skogsindustrin visades för 72 procent personer med vit hudfärg och 90 procent män. Annonsering efter kassapersonal till livsmedelsbutiker visades för 85 procent kvinnor. Annonser för tjänster inom taxibranschen visades för 75 procent personer med svart hudfärg.[76] Därutöver visar forskning att det föreligger en risk för att annonser till välbetalda tjänster oftare visas för män. Det beror på att män i större utsträckning vid tidigare tillfällen har tenderat att klicka på annonser för den typen av tjänster.[77]

Algoritmerna riskerar att reproducera redan etablerade mönster då algoritmerna exempelvis får tillgång till data rörande hur arbetsgivares befintliga personal ser ut eller vilka arbetssökande som arbetsgivaren – eller andra liknande arbetsgivare – tidigare har bedömt som intressanta kandidater att kalla till intervju. Det gör de genom att rikta annonserna eller föreslå kandidater från vissa grupper som är överrepresenterade – och andra underrepresenterade – inom en viss bransch eller bland de som arbetsgivaren eller andra liknande arbetsgivare tidigare har visat intresse för. Detta kan ha samband med kön, etnisk tillhörighet, ålder eller någon

76 Ali, Sapiezynski, Bogen, Korolova, Mislove och Rieke (2019).
77 Köchling och Wehner (2020).

annan diskrimineringsgrund.[78] Forskning visar att dessa metoder inte leder till att arbetsgivare rekryterar andra personer än de som väljs ut genom traditionella metoder.[79]

Urval bland jobbsökande

Urval baserat på historiska data om tidigare anställda kan bidra till diskriminering i framtiden

Företag och organisationer använder i allt större utsträckning beslutsverktyg baserade på AI och annat automatiserat beslutsfattande exempelvis för att göra ett urval bland ansökningar, eller för att bearbeta information från telefon- eller videointervjuer i syfte att välja ut vilka kandidater som ska kallas till anställningsintervjuer.[80]

Diskriminering i samband med urval baserade på AI och annat automatiserat beslutsfattande kan uppkomma som ett resultat av att de stora datamängder som algoritmerna tränas på är snedvridna.[81] Urvalsalgoritmer tränas exempelvis ofta på data som innehåller information om tidigare rekryteringsbeslut inom en organisation eller andra liknande organisationer samt information om nuvarande anställdas prestationer.

Urvalssystem som baseras på tillgänglig data över vilka som lyckas bra i en organisation eller i en viss bransch riskerar i sin tur att identifiera variabler eller kombinationer av variabler som är indirekt relaterade till diskrimineringsgrunderna och basera sitt urval på dessa.[82] Risken är överhängande att algoritmen kommer att identifiera variabler som är jämförda med män om befintlig och historisk statistik rörande vilka som har rekryterats pekar på att de flesta som har rekryterats eller bedömts

78 Kim och Scott (2018); Bogen (2019).
79 Horton (2017).
80 Mann och O'Neil (2016); van Esch, Black och Ferolie (2019).
81 Kim (2017).
82 Drage och Mackereth (2022); Cheng och Hackett (2021).

som högpresterande är män. Det finns exempelvis flera rapporterade fall där algoritmerna gör urval som gynnar män framför kvinnor, eftersom män är överrepresenterade bland dem som tidigare har anställts till vissa typer av tjänster eller som har gjort karriär inom vissa organisationer.[83] Denna typ av diskriminerande effekter kan uppkomma även om information på ett direkt sätt rör en eller flera av diskrimineringsgrunderna ur den data som algoritmerna tränas på.[84]

I en amerikansk studie försökte forskarna via maskininlärning utveckla en algoritm som skulle identifiera kön (man/kvinna) och hudfärg (vit/icke-vit) hos personer som i sina ansökningshandlingar aktivt avstått från att ange denna information.[85] Studien visade att algoritmen på ett korrekt sätt kunde identifiera kön och hudfärg i 95 procent av fallen efter att den hade tränats på ett omfattande datamaterial. Det finns därmed risker för att ojämlikheter och skillnader som präglar många organisationer och arbetsmarknaden i stort kommer att återspeglas i den data algoritmerna tränas på och reproducera dessa mönster.[86] Det beror på att befintliga arbetsmarknadsrelaterade data präglas av skillnader och ojämlikheter grundade på kön och etnisk tillhörighet.[87]

83 Yang (2021).
84 Williams, Brooks och Shmargad (2018); Goretzko och Finja Israel (2022).
85 Sajjadiani, Sojourner, Kammeyer-Mueller och Mykerezi (2019).
86 Yarger, Cobb Payton och Neupane (2019); Bogen (2019).
87 Fountain (2022); Hoffmann (2019).

Vidare diskuterar artiklar på området det som i organisationspsykologisk litteratur benämns person-culture fit, det vill säga i vilken utsträckning en person passar in i den befintliga organisationskulturen och delar de värderingar som präglar organisationen och de som arbetar där.[88] Denna typ av urval är en känd riskfaktor för diskriminering som har identifierats i tidigare rekryteringsforskning, då det tenderar att leda till så kallad likhetsrekrytering. Det innebär rekrytering av personer som liknar de som redan finns i organisationen.[89]

Urval utifrån bostadsort, frånvaro och examensår kan ge diskriminerande effekter

Diskriminerande effekter kan också uppkomma som ett mer direkt resultat av hur modellen är utvecklad och konstruerad, exempelvis med avseende på vilka variabler som algoritmen baserar sina förutsägelser på. Att inte bo på alltför stort avstånd från arbetet har exempelvis visat sig påverka hur länge en person stannar kvar på en tjänst. Samtidigt kan information om en viss bostadsort förknippas med vissa etniska grupper, vilket innebär att modellen gör vissa förutfattade antaganden om människor som bor i dessa områden. På liknande sätt kan inkluderingen av variabler som exempelvis frånvaro eller tjänstledighet, enligt forskning, missgynna kvinnor med omvårdnadsansvar och personer med funktionsnedsättningar.[90] Information om hur många år som gått sedan man avslutade sin utbildning eller omfattningen av arbetslivserfarenhet riskerar i sin tur att utesluta kandidater på grund av ålder. Risker för diskriminering kan också uppkomma enbart grundat på det faktum att systemen inte fungerar lika bra för grupper som har varit underrepresenterade i det datasystemet har blivit tränat på. System som bearbetar videoinspelade intervjuer för att föreslå kandidater presterar

88 Vandenberghe (1999).
89 Björklund, Bäckström, och Wolgast (2012); Knocke, Drejhammar, Gonäs och Isaksson (2003).
90 Yang (2021).

mer oförutsägbart och sämre för grupper som är underrepresenterade i den data de har tränats på. Det gäller exempelvis kvinnor ur minoritetsgrupper baserat på etnisk tillhörighet eller personer med vissa funktionsnedsättningar.[91]

Urval genom datorspelslikande mätningar kan missgynna vissa grupper

Ett relaterat problem handlar om att olika variabler använda som mått på prestation eller lämplighet kan vara utformade och framtagna på ett sätt som gör att vissa grupper gynnas medan andra grupper missgynnas. Mätningar av egenskaper som är tänkta att vara neutrala i relation till diskrimineringsgrunderna, såsom intelligens, motivation och olika personlighetsdrag, sker i själva verket i ett socialt och historiskt sammanhang som undergräver denna förmodade neutralitet.[92]

Det gäller till exempel olika former av datorspelsliknande mätningar av egenskaper och färdigheter. Syftet med metoderna är att mäta egenskaper, exempelvis problemlösningsförmåga och förmåga till beslutsfattande, men mätningen sker på ett sätt som gynnar vissa grupper och missgynnar andra. Studier visar att män tenderar att prestera bättre än kvinnor och unga bättre än äldre beroende på skillnader i vana att hantera och agera i datorspelsliknande sammanhang och situationer.[93]

91 Mehrabi, Morstatter, Saxena, Lerman, och Galstyan (2022); Koechling, Riazy, Wehner, och Simbeck (2020); Kelly-Lyth (2021); Varma, Dawkins och Chaudhuri (2022).

92 Duckworth, Quinn, Lynam, Loeber, Stouthamer-Loeber och Smith (2011); Mani, Mullainathan, Shafir och Zhao (2013); Henrich, Heine och Norenzayan (2010).

93 Melchers och Basch (2022).

En alltför stor tilltro till urvalssystem baserade på AI är en risk i sig

Det framkommer även i några forskningsartiklar att användare av urvalssystem baserade på AI och annat automatiserat beslutsfattande tolkar det utfall som systemen genererar som objektiva sanningar om kandidaternas relativa kvaliteter. Därmed följer de dessa rekommendationer mer okritiskt än vid urval genomförda av människor.[94] Denna så kallade automation bias innebär att de personer som är involverade i beslutsfattandet blir mindre uppmärksamma och mindre noggranna när ett algoritmbaserat hjälpmedel är inblandat. Det beror på att de tenderar att uppfatta algoritmer som rationella, vetenskapliga och värdeneutrala.[95]

En avgörande drivkraft för införandet av system som bygger på AI och annat automatiserat beslutsfattande är att hitta kostnadseffektiva och tidsbesparande sätt att gå igenom och bedöma ett stort antal ansökningar till utlysta tjänster.[96] Såväl forskare som kommersiella aktörer har framhållit fördelarna med AI och annat automatiserat beslutsfattande vid urvalsmetoder för att minska och förebygga diskriminering. Förhoppningen har varit att system baserade på AI och annat automatiserat beslutsfattande ska kunna åstadkomma ett urval som är opåverkat av fördomar, negativa attityder eller stereotyper, vilket ett stort antal studier har påvisat[97] rörande rekrytering och urval baserat på mänskligt beslutsfattande.[98]

Det finns flera artiklar som refererar till företag och systemutvecklare som hävdar att de system som de har utvecklat bidrar till en mer objektiv bedömning av de sökande till en viss tjänst, vilket därigenom motverkar diskriminering.[99]

94 Drage och Mackereth (2022); Busuioc (2021); Bolander (2019).
95 Peeters, Giest och Grimmelikhuijsen (2020).
96 Leicht-Deobald, Busch Schank, Weibel, Schafheitle, Wildhaber och Kasper (2019).
97 Se till exempel Quillian, Heath, Pager, Midtbøen, Fleischmann och Hexel (2019).
Drage och Mackereth (2022); Skanderson (2021).
98 Yarger, Cobb Payton och Neupane (2019); Savage och Bales (2017).
99 Suen, Chen och Lu (2019).

I en experimentell studie fann forskarna att bedömningar av hur väl lämpade arbetssökande var för en tjänst utifrån en videoinspelad intervju påverkades i signifikant större utsträckning av initiala intryck och utseende när bedömningarna gjordes av människor jämfört med när de byggde på AI och annat automatiserat beslutsfattande.[100]

En annan studie tillämpade AI och annat automatiserat beslutsfattande på inspelade intervjuer för att bedöma graden av motivation hos studenter som sökte ett fiktivt arbete.[101] Resultaten visade att de bedömningar som baserades på AI och annat automatiserat beslutsfattande stämde bättre överens med de sökandes självskattade motivation och var något mindre påverkade av den sökandes kön, jämfört med bedömningar av mänskliga rekryterare.

Liknande resultat konstateras i en studie om rekrytering till lärartjänster i en delstat i USA. I studien använde forskarna AI och annat automatiserat beslutsfattande för att systematiskt klassificera och bearbeta information om tidigare arbetslivserfarenheter i skriftliga ansökningar, för att därefter utifrån detta underlag föreslå lämpliga kandidater. Studien bygger på så kallad övervakad AI och annat automatiserat beslutsfattande, vilket innebär att i förväg bestämma vilka kategorier algoritmerna skulle lära sig. Här fann forskarna att det urval av kandidater som den algoritmbaserade modellen föreslog var mer neutralt i relation till kön (man/kvinna), hudfärg (vit/icke-vit) och ålder än vad urvalsbesluten hos arbetsgivarna i det aktuella skoldistriktet brukade vara.[102]

100 Yarger, Cobb Payton och Neupane (2019); Savage och Bales (2017).
101 Kappen och Naber (2021).
102 Sajjadiani, Sojourner, Kammeyer-Mueller och Mykerezi (2019).

AI och annat automatiserat beslutsfattande inom andra delar än rekrytering inom arbetslivsområdet

Det finns också en del artiklar som rör möjligheter vid användning av AI-system och risker för diskriminering med system baserade på AI och annat automatiserat beslutsfattande i relation till andra områden i arbetslivet. Det handlar dels om uppföljning och bedömning av anställdas prestationer, dels om beslutsfattande i befordransprocesser och i relation till löneutveckling.

Möjligheter och risker vid uppföljning och bedömning av anställdas prestationer

Företag och organisationer förlitar sig på algoritmbaserade modeller för att följa upp och utvärdera anställdas arbetsprestationer som underlag för arbetsledning och beslutsfattande.[103] System baserade på AI och annat automatiserat beslutsfattande har en omfattande kapacitet när det gäller att automatiserat och snabbt kunna bearbeta stora datamängder. Detta har medfört att sådana system i ökad utsträckning används för att samla in och bearbeta uppgifter om olika aspekter av anställdas prestationer och beteenden på arbetet och därmed öka effektiviteten.[104]

Det saknas dock artiklar som specifikt undersöker om dessa system kan användas för att förebygga diskriminering. Enstaka teoretiska resonemang lyfter fram möjligheten att system baserade på AI och annat automatiserat beslutsfattande inom detta område skulle kunna möjliggöra en mer objektiv och neutral utvärdering.[105] System baserade på AI och annat automatiserat beslutsfattande skulle därigenom kunna minska

103 Leicht-Deobald, Busch, Schank, Weibel, Schafheitle, Wildhaber och Kasper (2019); Peeters, Giest och Grimmelikhuijsen (2020).
104 Tambe, Cappelli och Yakubovich (2019).
105 Köchling och Wehner (2020); Charlwood och Guenole (2022).

effekterna av godtycke och fördomar vid prestationsutvärderingar, och på så sätt förebygga diskriminering.[106]

Forskningsartiklarna belyser en rad olika risker och negativa effekter. Merparten av dem handlar om risker och nackdelar i relation till anställdas integritet. De rör också negativa effekter av de ökade möjligheter till kontroll och övervakning som dessa system medför.[107] Forskare pekar exempelvis på system som registrerar och utvärderar anställdas rörelsemönster, emotionella uttryck i kundkontakter, aktivitet på sociala medier och internet, eller närvaro- och frånvarostatistik. Det finns även exempel på system som registrerar och bearbetar information om vad anställda i så kallade call-centers säger i kundsamtal, hur förare i transportföretag kör och exakta rörelsemönster hos monterings- och lagerarbetare.[108] Riskerna med denna utveckling återspeglas också i litteratur som visar att anställda reagerar negativt på dessa system. Införandet av dem associeras till negativa känslor, upplevd orättvisa, låg tillit, minskat engagemang i arbetet, ökad arbetsbelastning och lägre arbets- och livstillfredsställelse.[109]

De risker och negativa effekter denna litteratur lyfter fram inbegriper således inte diskriminering i sig. Utvecklingen mot ökad kontroll och övervakning sker dock huvudsakligen inom arbeten och i delar av ekonomin där det är känt att exempelvis unga och personer med viss etnisk tillhörighet är överrepresenterade. Mot den bakgrunden finns det därför en risk att den ökade användningen av artificiella system leder till

106 Parent-Rocheleau och Parker (2022).
107 Kellogg, Valentine och Christin (2020); Ajunwa (2018); Bales och Stone (2020); Kim och Bodie (2021); Tursunbayeva, Pagliari, Di Lauro och Antonelli (2022).
108 Parent-Rocheleau och Parker (2022); Leicht-Deobald, Busch, Schank, Weibel, Schafheitle.Wildhaber och Kasper (2019).
109 Lee (2018); Griesbach,Reich, Elliott-Negri och Milkman (2019); Keith, Harms och Tay (2019); Bucher, Fieseler och Lutz (2019); Hamilton och Davison (2022).

en försämrad arbetsmiljö och ökade integritetsintrång för dessa grupper jämfört med andra grupper på arbetsmarknaden.

Diskriminerande effekter kan uppstå i samband med befordran och lönesättning

System baserade på AI och annat automatiserat beslutsfattande används även i arbetslivet inom processer som relaterar till beslut om befordran och lönerevisioner. Algoritmbaserade modeller används exempelvis för att förutsäga framtida prestationer baserat på tillgänglig prestationsstatistik och nuvarande arbetsuppgifter. Det gäller även nyligen genomgångna utbildningar för att därigenom fatta beslut i relation till exempelvis befordringar och löneutveckling.[110] Även här öppnar således användandet av system baserade på AI och annat automatiserat beslutsfattande principiellt upp för möjligheter men också risker, i relation till diskriminering.

Det saknas artiklar som explicit undersöker möjligheten att förebygga diskriminering i samband med beslut om befordran eller i relation till lönesättning med hjälp av system baserade på AI och annat automatiserat beslutsfattande. Det finns dock enstaka artiklar som principiellt pekar på sådana möjligheter utifrån liknande resonemang som har förts i relation till andra områden. System baserade på AI och annat automatiserat beslutsfattande möjliggör enligt dessa artiklar ett mer neutralt och objektivt beslutsfattande som baseras på relevanta sakförhållanden, snarare än på ovidkommande eller godtyckliga faktorer som riskerar att resultera i diskriminering.[111]

110 Angrave, Charlwood, Kirkpatrick, Lawrence och Stuart (2016); Parent-Rocheleau och Parker (2022).

111 Parent-Rocheleau och Parker (2022); Köchling och Wehner (2020); Charlwood och Guenole (2022); Newman, Fast och Harmon (2020).

De risker för diskriminering som forskningsartiklarna lyfter fram rör till exempel beslut om befordran som baseras på förutsägelser från modeller tränade på ett datamaterial som reproducerar redan etablerade mönster av gruppbaserade skillnader. Dessa mönster kan kopplas till vilka som tidigare har blivit befordrade eller vilka som tidigare har bedömts som högpresterande anställda.[112] Prestationsutvärderande system kan få potentiellt diskriminerande effekter i samband med exempelvis lönebildning. Systemen riskerar att premiera och missgynna löneutvecklingen hos anställda om dessa utvärderingssystem inkluderar information om frånvaro (såsom sjukdom eller vård av barn) eller beteendeskillnader på gruppnivå i relation till någon av diskrimineringsgrunderna (exempelvis data rörande hur snabbt olika förare kör inom transportföretag). Detta på ett sätt som skulle kunna utgöra diskriminering.

Diskriminerande effekter vid beslut om befordran och löneutveckling gäller också om dessa baseras på modeller som utgår från data om hur de anställda har blivit bedömda av kunder i deras kontakter med dem.[113] Under sådana omständigheter riskerar nämligen negativa stereotyper eller attityder hos kunderna kring exempelvis personer med funktionsnedsättning, könsöverskridande identitet eller uttryck eller viss etnisk tillhörighet att påverka de data som lönesättningen baseras på.

Även på detta område finns forskning som visar att anställda uppfattar algoritmbaserat beslutsfattande som orättvist. Det upplevs som förminskande och baserat på faktorer som förvisso är kvantifierbara men som brister i relevans när de används och tolkas utanför sin kontext.[114] Andra studier pekar på att reaktionerna på att beslut i relation till ens egen person som har fattats av system baserade på AI och annat automatiserat beslutsfattande blir mer negativa ju viktigare beslut det rör

112 Köchling och Wehner (2020).
113 Wood, Graham, Lehdonvirta och Hjorth (2019).
114 Newman, Fast och Harmon (2020).

sig om.[115] På liknande sätt finns undersökningar som visar att chefer generellt gärna ser att system baserade på AI och annat automatiserat beslutsfattande används inom ramen för beslutsprocesser, men att mänskliga beslutsfattare ska ha det avgörande inflytandet över de slutliga besluten.[116]

Sammanfattande resultat

Forskningen pekar på både möjligheter vid användning av AI-system och risker med AI och annat automatiserat beslutsfattande i relation till olika processer i arbetslivet. Det rör sig exempelvis om tillämpningar i relation till rekrytering och urval, bedömning och utvärdering av anställdas prestationer och beslutsfattande i relation till lön och befordran. På samtliga dessa områden identifierar litteraturen risker för diskriminering som kan uppstå vid användning av AI och annat automatiserat beslutsfattande, men även möjligheter att använda tekniken för att minska eller förebygga diskriminering.

Motstridiga resultat om AI och diskriminering

Forskningsöversikten visar att det finns stora forskningsluckor om risker för diskriminering och AI och annat automatiserat beslutsfattande i arbetslivet. Det framkommer även att det vetenskapliga underlaget för många av de anspråk som görs på detta område är tämligen begränsat. Det beror på variation i hur system med AI och annat automatiserat beslutsfattande kan utformas, på vilken data de tränas och i vilket sammanhang de används.

115 Langer, König och Papathanasiou (2019).
116 Haesevoets, De Cremer, Dierckx och Van Hiel (2021).

Forskningen pekar därmed i många fall åt olika håll. Vissa artiklar visar på möjligheter med AI och annat automatiserat beslutsfattande när det gäller att förebygga diskriminering. Samtidigt lyfter andra artiklar fram riskerna för diskriminering när AI och annat automatiserat beslutsfattande används inom samma område och på liknande sätt.

Flera studier visar att system baserade på AI och annat automatiserat beslutsfattande kan minska risken för diskriminering i samband med urval av vilka kandidater som ska gå vidare i rekryteringsprocessen utifrån inkomna ansökningar. Det finns forskning som har pekat på att urvalet kan bli mer neutralt gällande variabler som kön, ålder och etnisk tillhörighet när det genomförs med hjälp av system baserade på AI och annat automatiserat beslutsfattande. Detta gäller i synnerhet om systemen aktivt har utformats för att säkerställa att urvalet är neutralt när det gäller dessa variabler.

Samtidigt visar forskning på det motsatta: att användning av AI och annat automatiserat beslutsfattande i samband med annonsering och urval också riskerar att leda till diskriminering. En stor andel av forskningsartiklarna pekar på risker för diskriminering utifrån hur system baserade på AI och annat automatiserat beslutsfattande på ett övergripande sätt fungerar och utvecklas. Det är emellertid oklart och återstår att undersöka hur omfattande dessa diskriminerande effekter i praktiken är i relation till de mest använda systemen på arbetsmarknaden. Detta gäller inte minst i en svensk kontext, där ingen publicerad artikel som har identifierats uppfyller villkoren för att inkluderas i forskningsöversikten.

Den spridningstakt som system baserade på AI och annat automatiserat beslutsfattande har haft i arbetslivet har varit långt större än motsvarande ökningstakt i vetenskaplig litteratur som visar att dessa system faktiskt har avsedda effekter. Den genomförda forskningsöversikten visar på en påtaglig kontrast mellan de anspråk på att kunna

säkerställa en diskrimineringsfri rekrytering som görs av en rad olika kommersiella aktörer på området, och det vetenskapliga stödet för att så verkligen är fallet. Detta betyder inte nödvändigtvis att de påståenden som systemutvecklarna gör om hur deras system fungerar är falska. Däremot hänvisar de till data och uppgifter från sina egna utvärderingar av systemen de marknadsför, och inte till systematiska undersökningar som har genomgått kritisk granskning av oberoende experter på området, vilket skulle ha krävts för vetenskapliga publikationer. Sådana omständigheter leder med nödvändighet till en lägre trovärdighet för dessa anspråk och påståenden, eftersom det finns ett tydligt ekonomiskt intresse att marknadsföra sina egna produkter som effektiva och lyfta fram dess fördelar.

Återspegling av ojämlikheter på arbetsmarknaden

Flera forskningsartiklar framför och visar att risker för diskriminering kan uppkomma eftersom den data som systemen tränas på återspeglar historiska och befintliga ojämlikheter och grupprelaterade skillnader i arbetslivet. Denna grundläggande omständighet riskerar exempelvis att leda till att systemen tolkar denna data som att etablerade mönster av över- och underrepresentation mellan grupper inom olika branscher eller på olika befattningsnivåer återspeglar skillnader i kompetens och lämplighet. Systemen riskerar då att fatta eller föreslå beslut som reproducerar dessa gruppskillnader.

Representationsbias leder till diskriminering

Flera artiklar lyfter även fram och påvisar risker för det som kan kallas representationsbias. Det innebär att när den data algoritmerna ska tränas på genereras kan det uppstå risker för diskriminering som en direkt effekt av vilka grupper som har varit underrepresenterade i träningsdatan. Detta beror då på att systemen har ett bristande underlag i relation till dessa grupper, vilket leder till att de tenderar att fungera mer oförutsägbart och sämre när de ska bedömas.

Risker för diskriminering kan också uppkomma baserat på vilka egenskaper eller variabler som ingår i det datamaterial som systemen har tränats på och hur dessa har mätts. Här pekar forskningen exempelvis på risker för såväl direkt som indirekt diskriminering om systemen baserar sina beslut på faktorer eller egenskaper som är direkt eller indirekt relaterad till någon av diskrimineringsgrunderna. Eftersom diskrimineringsgrunder som exempelvis kön och etnisk tillhörighet har konsekvenser för utfall och erfarenheter på en rad olika livsområden, framhåller forskning att det i praktiken är svårt att skapa system som helt kan förbise dessa variabler. Det finns exempelvis studier som visar att system baserade på AI och annat automatiserat beslutsfattande kan användas för att identifiera kön och hudfärg hos personer med stor träffsäkerhet, även i datamaterial där all direkt information om dessa faktorer har utelämnats.

Forskning på området visar även att vissa av de mätningar som system baserade på AI och annat automatiserat beslutsfattande utgår ifrån i sitt beslutsfattande i realiteten kan påverkas av ovidkommande faktorer eller genomföras via metoder som kan ha diskriminerande effekter. Det gäller exempelvis videoinspelade intervjuer som används för att bedöma personlighetsdrag och datorspelsliknande tester som används för att mäta problemlösningsförmåga.

Ytterligare ett exempel på hur diskriminering kan uppkomma i detta skede är om träningsdatan inkluderar variabler som i sig är påverkade av negativa stereotyper eller attityder kopplade till någon av diskrimineringsgrunderna. Detta kan exempelvis vara fallet om kundnöjdhet, det vill säga kundernas omdöme om den anställde, inkluderas i datamaterial som ska ligga till grund för beslut om anställning, befordran eller lön.

Automation bias – att okritiskt acceptera utfallet av systemen

Flera studier visar även att mänskliga beslutsfattare som agerar i relation till system baserade på AI och annat automatiserat beslutsfattande riskerar att bli mindre noggranna och kritiska. I stället tenderar de att uppvisa det som i litteraturen benämns automation bias, vilket innebär att de okritiskt accepterar utfallet av systemen som något neutralt, objektivt och värderingsfritt. Riskerna med detta är uppenbara, givet de flera möjliga vägar till diskriminerande utfall i sådana system som har diskuterats ovan. Risken är därmed att diskriminerande utfall som systemen genererar inte upptäcks och korrigeras om användarna inte är medvetna om att dessa system – precis som alla andra metoder för informationsbearbetning och beslutsfattande – har såväl för- som nackdelar och måste granskas kritiskt.

Forskare pekar också på att system baserade på AI och annat automatiserat beslutsfattande kan användas för att osynliggöra eller legitimera diskriminering. Här menar forskarna att arbetsgivare kan motivera och rättfärdiga utfall och beslut även om diskriminering har förekommit genom att hänvisa till utfallen från dessa system som icke-diskriminerande och sakliga.

AI-systemen kan leda till systematisk diskriminering

Den snabba spridningen av system baserade på AI och annat automatiserat beslutsfattande har inte minst motiverats av deras överlägsna förmåga att bearbeta stora mängder information snabbt, systematiskt och konsekvent. Samtidigt finns det i denna systematik och konsekvens en risk som har lyfts fram i några av de inkluderade artiklarna i forskningsöversikten, nämligen att dessa system – i den utsträckning de är diskriminerande – även kommer att åstadkomma diskriminerande utfall konsekvent och systematiskt.

En relevant aspekt i sammanhanget är att denna risk är i högsta grad relevant och problematisk, även om systemen skulle vara mindre diskriminerande än mänskliga beslutsfattare i de enskilda fallen. Detta beror på att det kommer att finnas en stor variation i hur diskrimineringen på grund av mänskliga beslutsfattare ser ut och vilka konsekvenser denna diskriminering får i enskilda fall. Sådan diskriminering har en föränderlighet som drivs av en rad omständigheter som har med såväl person- som situationsfaktorer att göra. Den diskriminering som sker riskerar emellertid att bli mer enhetlig, konsekvent och systematisk om ett begränsat antal system baserade på AI och annat automatiserat beslutsfattande, eller flera olika system som fungerar på liknande sätt, dominerar inom ett område. Denna risk handlar alltså inte om att system baserade på AI och annat automatiserat beslutsfattande nödvändigtvis skulle vara mer – eller ens lika – diskriminerande i enskilda fall. Risken utgörs i stället av att en ökad användning av sådana system skulle göra diskriminerande processer och utfall mer systematiska och konsekventa.

En professor i datavetenskap vid Princeton uttrycker denna farhåga på följande vis (vår översättning):

> Mänskliga beslutsfattare kan vara partiska, men det finns i alla fall en mångfald i partiskheten. Föreställ dig en framtid där alla arbetsgivare använder algoritmer för automatiserad screening av ansökningar som alla använder samma heuristik, och att arbetssökande som inte passerar dessa kontroller blir avvisade överallt.[117]

117 Citat i Benjamin (2019).

Kapitel 5 Arbetsgivares användning av AI och kunskap om risker för diskriminering

Detta kapitel redovisar en undersökning av i vilken utsträckning och i vilka sammanhang som arbetsgivare i Sverige använder sig av AI och annat automatiserat beslutsfattande i arbetslivet. Undersökningen bygger på telefonenkäter med 10 rekryteringsföretag och 100 arbetsgivare inom privat sektor. Kapitlet beskriver också vilken kunskap som finns bland arbetsgivare om vilka risker för diskriminering som kan finnas och vad de gör för att minimera riskerna.

Undersökningen har genomförts av Stefan Larsson och James White vid institutionen för teknik och samhälle vid Lunds Tekniska Högskola, Lunds universitet och Research Institute for Sustainable AI (RISAI) i samverkan med Novus. Kapitlet bygger på ett manus av Stefan Larsson och James White som DO har redigerat.

Syfte, frågeställningar och metod

Undersökningens syfte är att studera i vilken utsträckning och i vilka sammanhang som arbetsgivare i Sverige använder sig av vissa AI och annat automatiserat beslutsfattande i arbetslivet och identifiera vilka risker för diskriminering som finns inbyggda i dessa tekniska lösningar. Syftet är även att undersöka vilken kunskap som finns bland arbetsgivare om dessa risker och vad de gör för att minimera riskerna.

Kapitlet försöker besvara följande frågor:

- I vilken utsträckning och i vilka sammanhang använder sig arbetsgivare i Sverige av AI och annat automatiserat beslutsfattande i arbetslivet?
- Vilka risker för diskriminering finns inbyggda i dessa tekniska lösningar?
- Vilken kunskap om eventuella risker för diskriminering har användare av dessa tekniska lösningar?

Forskarna har i samverkan med Novus och i dialog med DO tagit fram en undersökningsdesign. Initialt genomförde forskarna några kortare pilotintervjuer med experter som på olika sätt hade relevant erfarenheter av frågor inom AI, arbetsliv, rekrytering och diskriminering. Experterna bidrog till förtydliganden i de två enkäter som användes i undersökningen.

Den baseras dels på telefonenkät med 10 av de största rekryterings-företagen i Sverige, dels på telefonenkät med 100 företag med fler än 100 anställda. Syftet med telefonenkäterna var att få en överblick över användningen av AI och annat automatiserat beslutsfattande vid rekrytering och arbetsledning. För mer information om metod se bilagorna 2 och 3.

Undersökningen påverkas av hur AI och automatiserat beslutsfattande förstås av respondenterna. Osäkerhet kring vad som avses med AI-begreppet medför två utmaningar, en metodologisk och en analytisk. Den metodologiska utmaningen ligger i att det inte går att förvänta sig att respondenter har en enhetlig förståelse för vad som är och inte är AI. Den analytiska utmaningen härrör från den stora bredden av tekniker och metoder som faller under paraplybegreppen AI och automatiserat beslutsfattande, och svårigheten med att veta när och var en AI-lösning används inom ett bredare system.

I stället för att definiera AI och annat automatiserat beslutsfattande på ett strikt, tekniskt sätt visades respondenterna en förklarande text som har utformats för att arbeta med, snarare än mot, deras förförståelse av dessa termer.

> Automatiserat beslutsfattande omfattar algoritmiska och automatiserade beslutsprocesser såväl med som utan artificiell intelligens. I begreppet inkluderar vi både helt automatiserat beslutsfattande och automatiserade processer som används som beslutstöd.

Detta pragmatiska tillvägagångssätt accepterar att det finns variationer i förståelsen av AI. Respondenterna tillfrågades om användning av AI och annat automatiserat beslutsfattande på flera sätt, inklusive om användningen av metoder och verktyg som förväntas innehålla olika grad av automation och AI, för att därigenom tydligare fånga hur respondenterna förstår begreppen. Den metodologiska utmaningen kringgås därmed och görs till en analys, som samtidigt möjliggör en undersökning av utbredning.

Detta kapitel refererar generellt till annat automatiserat beslutsfattande när distinktionen mellan de olika nyanser som beskrivs ovan inte är viktiga. Ett betydelsefullt särfall gäller automatiserat beslutsstöd, det vill säga när det automatiserade beslutsfattandet inkluderar en människa som tar beslut baserat på information som har samlats in eller sammanställts via något typ av automatiserad metod. Automatiserat beslutsstöd refereras specifikt när det utgör en viktigt poäng för den analys som görs.

Metodmässigt blir det därmed inte alltid tydligt vad respondenten åsyftar när hen svarar på intervjufrågorna i denna undersökning. Det syns inte minst i skillnaden mellan svar på direkta frågor om användning av AI jämfört med mer specifika svar på frågor om användning av verktyg och specifika praktiker. I det förstnämnda fallet anger respondenterna överlag en låg AI-användning, medan i det sistnämnda en relativt hög.

Rekryteringsföretags och arbetsgivares användning av AI och annat automatiserat beslutsfattande: omfattning och sammanhang

Detta avsnitt redogör för hur arbetsgivare använder sig av AI och annat automatiserat beslutsfattande i samband med rekrytering och arbetsledning.

De huvudsakliga resultaten är:

- Verktyg baserade på AI och annat automatiserat beslutsfattande används i högre utsträckning än vad arbetsgivarna är medvetna om eller upplever.
- Stora arbetsgivare har en betydande sekundär användning av AI, det vill säga de använder lösningar som har förpackats i tjänster som används av företaget.
- Stora arbetsgivare lägger ut rekrytering på entreprenad, vilket sannolikt medför en stor användning av automatiserade beslutsstöd.

Fler arbetsgivare använder AI och annat automatiserat beslutsfattande än vad de själva tror

I den första och andra frågan i båda telefonenkäterna ombads respondenterna att utvärdera i vilken utsträckning AI eller automatiserat beslutsfattande används inom deras organisationer.

Till rekryteringsföretagen riktades frågor om förekomsten av dessa tekniker inom tjänster som erbjuds kunder. Totalt 8 av 10 svarade att användningen var ganska låg eller mycket låg (tabell 1). Detsamma gällde för användningen av AI, där 8 av 10 svarade att användningen var ganska låg eller mycket låg. En majoritet, 8 av 10, anger att de får in uppgifter om lämpliga kandidater via sociala medieplattformar, till exempel Facebook, LinkedIn, Instagram, TikTok, Twitter med flera.

Tabell 1: Fråga 4 till rekryteringsföretag. Hur får ni in olika uppgifter om lämpliga kandidater?

Flera svar möjliga. Totalt antal intervjuer: 10.

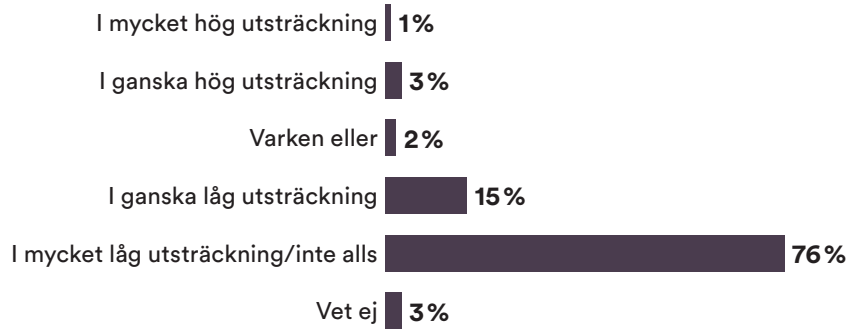
Svarsalternativ till fråga 4	Antal ja-svar
Vi testar kandidater genom spel, IQ-tester eller personlighetstester	9
Vi ber dem spela in en video med svar, som vi analyserar	1
Från egna webb-formulär	7
Webbanvändningsdata inklusive cookies från den sökande på er egen hemsida	3
Från sociala medieprofiler (till exempel Facebook, LinkedIn, Instagram, TikTok, Twitter med flera)	8
Olika befintliga plattformar, exempelvis på Arbetsförmedlingen, Monster med flera	10
Från datamäklare som samlat på sig profiler genom olika sorters datainsamling, och erbjuder till försäljning	1
Annat, nämligen	4
Vet ej	0

För arbetsgivarföretagen gällde frågorna användningen av AI eller automatiserat beslutsfattande inom personal- och arbetsledning. Här var den rapporterade användningen också låg, 87 procent svarade att olika former av automatiserat beslutsfattande eller automatiserat beslutsstöd vid rekrytering var ganska låg eller mycket låg (figur 1).

Figur 1: Fråga 1 till arbetsgivare. I vilken utsträckning inkluderar de system ni använder för att hantera rekrytering olika former av automatiserat beslutsfattande eller automatiserat beslutsstöd?
Antal intervjuer: 100.

Totalt 91 procent svarade att deras användning av automatiserat beslutsfattande/beslutsstöd i arbetsledning var ganska låg eller mycket låg (figur 2). Detta ska dock ses i ljuset av den högre frekvens som anges när specifika verktyg och tjänster som används efterfrågas.

Figur 2: Fråga 2 till arbetsgivare. I vilken utsträckning använder ni er av olika former av automatiserat beslutsfattande eller automatiserat beslutsstöd i arbetsledning som har med medarbetare att göra?
Antal intervjuer: 100.



Det går inte att anta att detta är en helt korrekt representation av den faktiska användningen av dessa tekniker hos svenska arbetsgivare, med tanke på att respondenterna förväntas ha något olika uppfattningar om vad som utgör AI och annat automatiserat beslutsfattande. Vad dessa resultat erbjuder är snarare en översikt över deras upplevda användning. Den rapporterade användningen var mycket högre när specifika applikationer och tjänster efterfrågades. Totalt 7 av 10 av rekryteringsföretagen svarade att de matchar kandidater med riktad annonsering till kandidaternas professionella profil (tabell 2, fråga 5). Det tyder på att respondenterna trots allt använder ett slags AI-funktionalitet med stort inslag av automation i sin rekryteringsprocess, då en rad studier visar att det är vanligt i jobbmatchningssammanhang att använda AI och annat automatiserat beslutsfattande för att automatiserat matcha arbetssökande med jobb, exempelvis via LinkedIn.

Tabell 2: Fråga 5 till rekryteringsföretag. Använder ni AI-system eller andra verktyg för att ...?

Flera svar möjliga. Totalt antal intervjuer: 10.

Svarsalternativ till fråga 5	Antal ja-svar
Matcha aktiva arbetssökande med jobb utifrån deras professionella profil	2
Matcha kandidater (både aktiva och passiva) med riktad annonsering mot deras professionella profil	7
Sortera eller ranka sökande för arbetsgivare	3
Annat, nämligen	2
Vet ej	0
Använder ej AI-system eller andra verktyg för automatiserat beslutsfattande	3

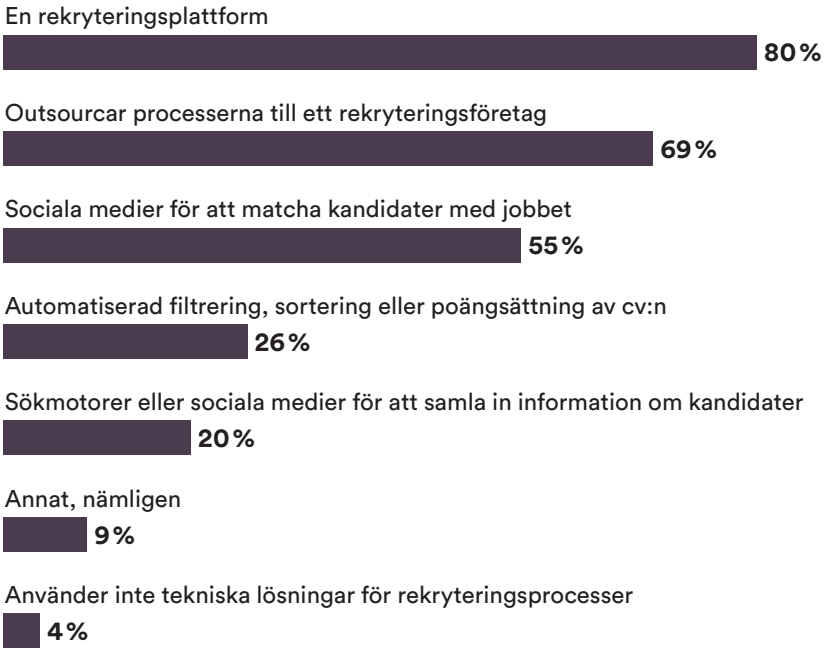
En majoritet av rekryteringsbolagen (9 av 10) testar kandidater genom spel, IQ-tester eller personlighetstester. Alla 10 anger att de använder sig av olika befintliga jobbplattformar, till exempel Arbetsförmedlingen eller Monster. Jobbplattformar erbjuder sökfunktioner som kan rangordna

träffar mot antingen jobbannonser (för jobbsökande) eller jobbsökandes cv:n (för rekryterare) med varierande grad av matchning genom automatiserade metoder.

På samma sätt rapporterade respondenterna i den större arbetsgivar-undersökningen en hög användning av tekniska lösningar i sina anställningsmetoder, inklusive rekryteringsplattformar (80 procent) – till exempel ReachMee, Varbi, LinkedIn Recruiter, Teamtailor – och sociala medier (55 procent) – till exempel Facebook, LinkedIn. Se figur 3.

Figur 3: Fråga 3 till arbetsgivare. Det finns olika tekniska lösningar som ibland används i rekryteringsprocesser. I era rekryteringsprocesser, använder ni er av något av följande?

Flera svar är möjliga. Antal intervjuer: 100.



Totalt 19 procent angav att de använder automatiserade former av filtrering, sortering eller poängsättning för att underlätta utvärdering av anställda, och 21 procent angav att de samlar in kvantitativa data om anställdas prestationer till exempel med hjälp av program som Deel, Papaya, Microsoft Teams employee monitoring. Se figur 4.

Figur 4: Fråga 4 till arbetsgivare. Det finns olika tekniska lösningar som ibland används för personalledning. I er personalledning, gör ni något av följande?

Flera svar är möjliga. Antal intervjuer: 100.

Det är tydligt att automatiserade beslutsstöd ofta används när det gäller rekryteringsföretagen. Omfattningen av automatisering eller användning av AI inom beslutsfattandet framgår inte tydligt av den nu genomförda undersökningen. Användningen hos stora arbetsgivarna är inte entydig. Rekryteringshanteringsplattformar, som ReachMee och Varbi, är integrerade lösningar som har utvecklats för att effektivisera platsannonsering och hantering av kandidater. De marknadsförs inte som AI-drivna lösningar, men använder former av automatiserad analys för beslutsstöd. De kan inkludera former av automatiserat beslutsfattande enligt den definition som används i denna undersökning. LinkedIn Recruiter har öppet diskuterat sin användning av AI för att matcha kandidater till jobb och hjälpa till att identifiera potentiella rekryterings-möjligheter.[118]

Medvetenheten om den vardagliga användningen av AI och annat automatiserat beslutsfattande på den svenska arbetsplatsen är låg, vilket dessa resultat sammantaget tyder på. Det kan förklara varför man inte uppmärksammar den roll som dessa tekniker spelar i vardagliga verktyg. Varje gång man utför en webbsökning (till exempel med Google) används ett automatiserat rekommendationssystem. Varje gång man scrollar igenom någons sociala medieflöde (till exempel på Facebook) är det ett AI-system som avgör vad man ser. Varje annons på dessa webbplatser och digitala plattformar har valts ut för oss av en komplex AI-modell, samtidigt som detta blir någonting vardagligt som vi inte lägger märke till. På samma sätt kan användningen av AI och annat automatiserat beslutsfattande på arbetsplatsen förväntas blekna i bakgrunden och bli vardaglig.[119]

118 Se LinkedIn (9 oktober 2018), Forbes (2 juli 2020).
119 Se Lomborg (2022) och Kaun (2022).

Stora arbetsgivares har betydande sekundäranvändning av AI-system

Man måste börja med att reda ut olika användningsformer för att kunna fördjupa frågan om omfattningen av AI och användning av annat automatiserat beslutsfattande hos svenska arbetsgivare. Detta kräver viss bakgrundsinformation om hur AI-baserade lösningar för närvarande kommer ut på marknaden.

Efterfrågan på kompetens inom AI-utveckling är hög. De typer av AI-system som för närvarande är populära hos teknikföretag (exempelvis data- och beräkningsmaximering) kräver höga nivåer av utbildning och erfarenhet, medan grunderna i maskininlärning är tillgängliga för alla med en dator och en outvecklad förståelse för programmeringsspråket Python. Som ett resultat har expertis inom AI-utveckling samlats inom Big Tech-företag (till exempel Apple, Google, Microsoft) och de specialiserade AI-innovationsföretagen (till exempel OpenAI). AI upplevs troligen som en uppsättning funktioner inom en programvara eller digital plattform för de allra flesta företag, och särskilt för små och medelstora företag. Dessa företag kan därmed få en begränsad förståelse för när och hur AI används i och som en del av dessa verktyg, och även de som har sådan expertis kan ha få möjligheter att anpassa AI-modeller för sina egna ändamål. Samma krafter är i spel även om situationen inte är lika uttalad för annat automatiserat beslutsfattande.

Denna dynamik innebär att det är analytiskt relevant att skilja primär och sekundär användning av AI och annat automatiserat beslutsfattande. Den primära användningen avser lösningar som har utvecklats i betydande utsträckning av de företag som använder dem, och den sekundära som användandet av lösningar som har förpackats i tjänster som används av företaget.

Två direkta utmaningar finns här. Den första utmaningen är att det finns en viss överlappning mellan de två kategorierna och att många av de alternativ som har erbjudits respondenterna i undersökningen inte entydigt är det ena eller det andra. Utan att uttryckligen fråga om systemen i fråga har utvecklats internt är det bästa som kan erbjudas en uppskattning av deras användning. Den andra utmaningen är att sekundär användning omfattar en mängd olika produkter och tjänster. En användbar distinktion att göra är mellan de som är lättillgängliga (till exempel Google-sökning) och de som medvetet har förvärvats och därför kan förväntas komma med stöd (till exempel Varbi). En annan distinktion kan göras mellan de produkter som inkluderar AI som en tillfällig funktion (till exempel Microsoft Word) och de vars funktionsuppsättning till stor del är uppbyggd kring AI-modeller (till exempel LinkedIn Recruiter). Med sistnämnda följer därmed potentiellt de risker för diskriminering som tecknas i denna rapport kring hur AI-modeller tränas, men med tillägget att det är svårt för någon extern aktör att granska storskaliga system. De flesta sekundära användningsområden kommer dock att finnas någonstans mellan dessa ytterligheter.

På frågan om tekniska lösningar i rekryteringsprocesser svarade 26 procent av de stora arbetsgivarna att de använde automatiserad cv-filtrering, sortering eller poängsättning (figur 3). Även om detta alternativ var avsett för att mäta primär användning är det mer sannolikt att det är en blandning av både primär och sekundär användning. Mycket vanligare är att företag använder en rekryteringsplattform (80 procent av respondenterna anger att de gör det, se figur 3), där denna typ av kandidatsortering kan integreras. Totalt 55 procent av de svarande rapporterade att de använde sociala medier (inklusive LinkedIn) för att matcha kandidater med jobb, och 20 procent rapporterade att de använde sökmotorer eller sociala medier för att hitta bakgrundsinformation om potentiella kandidater, se figur 3. Dessa resultat indikerar att sekundär användning av AI och annat automatiserat beslutsfattande är betydligt

högre än primär användning, och att dessa användningar oftast är bundna till en mjukvarulösning för vilken ett visst stöd finns tillgängligt.

Det kan inte göras ett definitivt mått på i vilken utsträckning AI-drivna lösningar används i rekryteringsprocesser utan att i enkätresultaten helt kunna skilja LinkedIn från andra rekryteringsplattformar (och med många företag som verkar använda en blandning av rekryteringsmetoder). Med detta sagt pekar resultaten på att användning av LinkedIn är vanligt i dagsläget. En tolkning av resultaten är att cirka två tredjedelar av företagen som använder en rekryteringsplattform använder just den plattformen i synnerhet. Det betyder att en stor del av bedömningen av eventuella risker för diskriminering genom användning av AI och annat automatiserat beslutsfattande vid rekrytering även leder till en bedömning av hur LinkedIn hanterar de omvittnade risker som finns med AI och annat automatiserat beslutsfattande generellt. Denna undersökning erbjuder inte någon direkt inblick i detta, mer än via de effekter som syns i de två enkäterna med rekryterare och arbetsgivarföretag. Detta talar för att mer uppmärksamhet bör riktas mot den roll de storskaliga digitala plattformarna spelar för diskrimineringsfrågorna i allmänhet, och vilka åtgärder de vidtar för att minska riskerna för diskriminering i synnerhet

Medan endast 4 procent av de stora arbetsgivarna angav att de inte använde en teknisk lösning för rekrytering (det vill säga deras anställning genomfördes på mer manuella och interpersonella sätt, se figur 3) framkom ett annat resultat inom personalledning. Här rapporterade 76 procent av de svarande att de inte använde beslutsfattande eller automatiserat beslutsstöd i arbetsledning som har med medarbetare att göra, se figur 2. Endast 4 procent angav att de använder AI och annat automatiserat beslutsfattande för att fastställa skäl för omplacering eller uppsägning, se figur 4. Ett större antal använder automatiserad schemaläggning av skift eller uppgiftstilldelning (14 procent), men våra resultat tydliggör inte om dessa processer tar hänsyn till personuppgifter eller prestationsmått. Totalt 12 procent använder automatiserad

poängsättning för medarbetarutvärderingar, se figur 4. Dessa svar överlappade varandra i hög utsträckning. Företag som använder automatiserat beslutsfattande i ett område gör sannolikt också det i flera andra.

Uppgifter om förekomsten av plattformar var mer tvetydiga inom personalledning än vid rekrytering. Totalt 21 procent av de stora arbetsgivarna angav att de samlar in kvantitativa uppgifter om anställda och deras prestationer, se figur 4. Det är möjligt att bekräftande svar även betyder att datainsamlingen görs på mer skräddarsydda sätt, medan många sannolikt kommer att göra det med hjälp av program som Microsoft Teams, med inbyggda möjligheter till övervakning av arbetstagare.

Rekrytering på entreprenad är vanligt bland stora arbetsgivare

Det är mycket vanligt att lägga ut rekrytering på entreprenad, så kallad outsourcing. Totalt uppgav 69 procent av de stora arbetsgivare som har genomfört undersökningen att de lägger ut någon del av sina rekryteringsprocesser till ett rekryteringsföretag, se figur 4. En majoritet om 80 procent anger att de använder en rekryteringsplattform och 55 procent att de använder sociala medier i sina rekryteringsprocesser. Det är möjligt att vissa respondenter tolkade outsourcing som den ena eller den andra av dessa tjänster. Det framgår inte heller av uppgifterna i vilken utsträckning och i vilken kombination respondenterna använder dessa olika tjänster.

På frågan hur rekryteringsbolagen får fram uppgifter om arbetssökande svarade samtliga att de använder olika befintliga plattformar, exempelvis Arbetsförmedlingen och Monster. Intressant i sammanhanget är att 8 av 10 använder sociala medieprofiler, vilket tyder på en externt beroende datainsamling. Sammantaget med att de flesta rekryterare (7 av 10, tabell 2) också anger att de använder sig av riktad annonsering för att matcha kandidater med riktad annonsering mot deras professionella profil visar på beroendet av externa, i hög grad automatiserade digitala plattformar.

Användningen av automatiserade beslutsstöd är därmed troligen hög bland rekryterare. I likhet med stora arbetsgivare kommer dessa sannolikt att vara en blandning av primära och sekundära användningsområden.

Risker för diskriminering finns inbyggda i tekniska lösningar

Detta avsnitt redogör för risker för diskriminering vid tillämpning av AI och annat automatiserat beslutsfattande i samband med rekrytering och arbetsledning. Det knyter an till frågan om vilka risker för diskriminering som finns inbyggda i tekniska lösningar.

De huvudsakliga resultaten är:

- Orättvisor kan automatiseras vid användningen av AI och annat automatiserat beslutsstöd på grund av snedvriden data.
- AI-modeller riskerar att reproducera strukturella fördomar.
- Bristande transparens gör det svårt att identifiera risker för diskriminering.
- Det finns i praktiken ingen entydig ansvarsfördelning gällande vem som bär ansvar för risker för diskriminering.

Partiskhet i data och automatiserade orättvisor

AI och annat automatiserat beslutsfattande marknadsförs ibland som ett opartiskt och neutralt alternativ till mänskligt omdöme. Argumentet är att eftersom maskiner inte har uttryckliga eller undermedvetna preferenser för individuella egenskaper eller eftersom de inte bär på fördomar mot grupper av människor, kan de inte diskriminera. Relaterat till detta är påståendet att eftersom maskiner inte blir trötta eller missnöjda med sitt arbete, kan de inte göra samma misstag eller fel i bedömningen som människor ofta gör. Det finns nu en betydande mängd studier som visar att detta argument inte riktigt stämmer. Anledningen till detta är enkel.

Systemen kan inte undkomma mänskliga fördomar och ojämlikheter eftersom de utvecklas och även tränas av människor och på data som representerar mänskligt språk och samhälle. Det innebär att AI och annat automatiserat beslutsfattande alltid medför risk för diskriminering, som därmed behöver granskas och hanteras så att de kostnadseffektiva och precisa matchningseffekterna som ofta framhävs kan nås utan allvarliga och oönskade effekter.

Komplexa rekommendationssystem kan fånga upp partiskhet eller så kallad bias som är djupt inbäddad i historiska och skeva samhällsstrukturer. Det finns exempel på forskning som har kunnat visa att sökresultat för webbplatser, bilder och geografiska platser har en systematisk brist på trovärdig och tillförlitlig information om kvinnor och människor med mörk hudfärg. Detta förstärker rasistiska och förtryckande stereotyper och de strukturella dominansförhållanden som ligger till grund för dem.[120]

Språkmodeller är tränade på stora volymer av ostrukturerad data och omfattar statistiska relationer mellan ord och grupper av ord inom neurala nätverk som kan vara mycket svåra att granska. Så om träningsdata för dessa system innehåller mänskliga fördomar, speglar systemen dessa fördomar.[121] Resultatet av stora språkmodeller kommer att innehålla viss inkonsekvens och oförutsägbarhet även om fördomsfulla uttalanden helt och effektivt tas bort från träningsdata. Det medföljer en risk för diskriminering.

120 Noble (2018).

121 För en rättssociologisk analys av utmaningarna med AI-system som speglar mänskliga sociala strukturer, se Larsson (2019).

Dessa typer av problem gäller också så kallad bildanalys eller datorseende inriktat mot människor, till exempel ansiktsigenkänningssystem. AI-modellerna i dessa system tränas på datauppsättningar med märkta bilder eller videor. Dessa data har ofta producerats av tusentals individer som bjuder på små, ”on-demand”-jobb som publiceras på ”crowdsourcing”-arbetsmarknader. Granskningar av dataset som har producerats på detta sätt avslöjar fortfarande problem och motstridigheter även om detta gör det möjligt för flera personer att dubbelkontrollera en enskild datapunkt.[122] Dessa problem och motstridigheter uppstår på grund av ofta upprepade fel och snedvridningar i taggningen till följd av den ojämna fördelningen av bilder i datauppsättningen. En studie av ansiktsigenkänningssystem fann att bristen på representation kopplad till etnisk mångfald inom träningsdata innebar att de resulterande modellerna presterade sämre på kvinnor och människor med mörk hudfärg.[123] Dessa resultat förvärrades för kvinnor med mörk hudfärg.

Det är långt ifrån klart att sådana problem med AI-driven bilddetektering kan lösas genom att helt enkelt fixa data. Inga tekniska framsteg kommer att få riskerna för diskriminering att försvinna helt. Det finns praktiska gränser, både när det gäller tid och pengar, för hur mycket utveckling som kommer att ingå i ett system innan det anses vara redo att tas i bruk. Detta kompliceras av systemens ökade komplexitet och av den grundläggande svårigheten att förmedla hur vissa typer av AI-system tränas. Dessutom finns det gränser för kunskap i sig, så att växande kunskap ofta åtföljs av en växande medvetenhet om osäkerhet, utanför vilken det verkligt ovetbara alltid kommer att bestå.[124] En viss ödmjukhet när det gäller beteendet och effekterna av dessa system är motiverad.

122 Crawford (2021).
123 Buolamwini och Gebru (2018).
124 White och Lidskog (2022).

AI-modeller riskerar att reproducera strukturella fördomar

AI och annat automatiserat beslutsfattande kan både användas för att förebygga diskriminering i rekrytering och leda till risker för diskriminering.

De fördelar som arbetsgivare och rekryterare vill nå har ofta med både kostnadseffektivitet och träffsäkerhet i matchning av jobb med kandidat att göra. LinkedIn, en stor rekryteringsplattform, har beskrivit hur deras AI-system, som automatiskt kan matcha kandidater med jobb, kraftigt har ökat effektiviteten i anställningsbeslut.[125]

Riskerna för diskriminering när AI-system och annat automatiserat beslutsfattande används vid rekrytering eller personalledning ligger exempelvis i hur automatisk screening och urval av sökande går till.

När en människa utan hjälp av automatiserade system gör en bedömning om en potentiell (eller nuvarande) anställd kan en medveten eller omedveten partiskhet påverka dem utan att de tror sig diskriminera på grundval av en eller flera av de sju diskrimineringsgrunderna. Under anställningen kan namnet på en kandidat implicit associera ett kön för bedömaren, eller etnisk tillhörighet eller till och med ålder. Detta kan i sin tur påverka hur mer relevanta uppgifter, såsom utbildning, erfarenhet och språkkunskaper uppfattas. Det kan i sin tur leda till ett beslut som riskerar att bli diskriminerande. Man kan tänka sig att sådana risker i mänsklig bias kan minskas genom att automatisering används som stöd för beslutsfattandet. Det finns också verktyg som till exempel hjälper rekryterare att använda mindre könspartiskt språkbruk i jobbannonser.

125 LinkedIn (den 9 oktober 2018), Forbes (den 2 juli 2020).

AI-system som använder maskininlärning erbjuder dock vissa särskilda utmaningar i relation till diskrimineringsrisker, oavsett om de helt och hållet förlitar sig på att fatta beslut eller helt enkelt används som stöd för dem. Här specificerar algoritmen inte vilka data som är relevanta och hur relevanta de är, utan snarare hur maskinen ska lära sig att väga vikten av data på egen hand. Slutsatserna som sådana system når kan vara överraskande. Kraften hos dessa modeller kommer från deras förmåga att känna igen korrelationer som människor inte kan, även i de fall data om kön och socioekonomisk status kan uteslutas från träningen av ett AI-system. Att utesluta kandidatens namn kan vara uppenbart, men att också utesluta deras fritidsaktiviteter kan vara att utelämna användbar information.

För att ta fram ett ändamålsenligt AI-system måste man hitta en balans mellan att styra modellen mot de data som anses vara mest relevanta, samtidigt som man inte överspecificerar hur sambanden förstås av människor. Risken i det senare är att man neutraliserar modellens förutsägbara förmågor för det den är tänkt att finna. Användningen av data som inte är taggad (annoterad) och organiserad i en förutbestämd struktur för så kallade oövervakade maskininlärningstekniker löper troligen ännu större risk att reproducera befintliga fördomar och ojämlikheter än användandet av data som är organiserad för övervakad maskininlärning. Det kan till exempel handla om att träna ett system på text från ett stort antal cv:n och liknande dokument utan betydande vägledning från utvecklaren om vad dessa data betyder. Modellen kan sedan välja lämpliga kandidater på grundval av önskade egenskaper i relation till jobbannonser. Här är det inte bara information om namn och plats som kan användas för att skilja mellan kandidater, utan också språkanvändningen i sig. Stavning, grammatik och ordanvändning korrelerar inte bara med uppfostran och utbildning utan kan också vara en markör för kön och etnisk tillhörighet.

Medan fokus i detta avsnitt har legat på textanalysens möjliga diskriminerande effekter, utgör så kallat datorseende, det vill säga ansikts- och affektigenkänning, ett annat potentiellt riskområde. Användningen av datorseende på den svenska arbetsplatsen verkar för närvarande vara låg. De data som krävs för att använda sig av detta inom rekrytering genereras inte i systematisk skala. Endast en av de tio rekryteringsbyråerna rapporterade att de använde sig av sökandevideor, se tabell 2, och endast 10 procent av de stora arbetsgivarna rapporterade att de hade tillgång till uppgifter relaterade till arbetssökandes utseende. Man kan anta att dessa bilder och videor inte analyseras av AI-system. Användningen av datorseende inom personalledning kommer dock sannolikt att bli vanligare i framtiden. Enligt undersökningen samlar 21 procent av de stora arbetsgivarna för närvarande in data om anställdas prestationer, se figur 4. Även om inte alla dessa kommer att använda webbkameror eller säkerhetskameror för att övervaka sin personal kan antalet som gör det förväntas växa. Integreringen av övervakningsfunktioner i vanlig programvara (till exempel Microsoft Teams eller Microsofts MyAnalytics[126]) och normaliseringen av distansarbete under covid-19-pandemin kommer sannolikt att tyda på en ökning av både utbud och efterfrågan på sådana verktyg. Diskrimineringsrisker med datorseende som används för att mäta ögonkontakt och ansiktsuttryck i video som används vid jobbansökningar diskuteras bland annat i relation till funktionsnedsättning, kön och hudfärg i amerikansk forskningslitteratur.[127]

126 Lomborg (2022).

127 För en analys utifrån en amerikansk rättslig kontext om risker för AI-relaterad diskriminering av personer med funktionsnedsättning, se Moss (2020); i relation till hudfärg och kön, se Buolamwini och Gebru (2018).

Riskerna kan även uppstå på andra områden. Totalt 9 av 10 respondenter i rekryteringsundersökningen angav att de använder spel, IQ-test eller personlighetstest för att bedöma arbetssökandes lämplighet, se tabell 2. Dessa former av bedömning innebär mentala och fysiska uppgifter som kanske inte tar hänsyn till en sökandes individuella behov som beror på funktionsnedsättning. På samma sätt är det på arbetsledningsområdet inte klart hur väl anpassade resultatmått är till behov och förmågor hos personer med funktionsnedsättning. I korthet, att mäta alla på samma sätt, med samma system, kan riskera att diskriminera personer med funktionsnedsättning.

Det finns även en kritik mot de mer tekniskt orienterade sätten att försöka fixa så kallad bias i kategorier som kön, etnisk tillhörighet och funktionsnedsättning, det vill säga att de inte är så enhetliga som de ofta uttrycks vara.[128] Kategoriseringarna innehåller stora variationer, vilket inte minst påpekas för funktionsnedsättning. Kategorin funktionsnedsättning låter sig inte enkelt fångas som dataklassificering och blir därmed extra utmanande även för de som försöker skapa rättvisare AI-system genom datoriserade metoder.[129]

Omvänt, mänsklig partiskhet och diskriminering har alltid funnits på arbetsplatsen. Många är fortfarande optimistiska över att AI och annat automatiserat beslutsfattande kommer att minska ojämlik behandling, även om det är uppenbart att de inte kommer att utrota dess förekomst. För det första fungerar mänsklig partiskhet och datordriven eller automatiserad partiskhet utifrån olika omfattning och skala. Mänsklig partiskhet tar sig många uttryck och varje individs fördomar kommer att ha en relativt begränsad effekt. Detsamma gäller inte för ett AI-system.[130] Eftersom AI-modeller tränas på stora datamängder

128 Se exempelvis Myers West med flera (2019).
129 Trewin med flera (2019).
130 Benjamin (2019).

kommer de att riskera att vara mer benägna att förstärka ett slags statistiskt medelvärde och riskera att reproducera vanligare förekommande, strukturella fördomar. På grund av den utbredda användningen av samma AI-system kan man inte utesluta att alla fördomar de har också kommer att vara utbredda, det vill säga riskera att diskriminera i stor omfattning.[131]

För det andra kan de fördomar som finns vara svåra att upptäcka. Komplexiteten i AI-system och vissa varianter av automatiserat beslutsfattande och tolkningsproblemet med tränade AI-modeller innebär att enbart programvarugranskningar inte nödvändigtvis är tillräckliga för att upptäcka diskriminering, vilket också gäller för ett ensidigt fokus på träningsdata.[132] Noggrann testning av distributionen av systemens resultat kommer också att krävas, vilket kan begränsas av tillgången på dataset att testa mot. Bara 1 procent av stora arbetsgivare (figur 5) och inga rekryteringsföretag angav att de har tillgång till data om nuvarande eller potentiella anställdas etniska tillhörighet. Ett vanligt antagande är att om systemet inte har tillgång till dessa uppgifter kan individer inte diskrimineras på den grunden, vare sig av en människa eller en maskin. Uteslutandet av denna information undanröjer dock inte risker för diskriminering som beskrivs ovan och kan förhindra eller undergräva inrättandet av effektiva kontroller eller motåtgärder. [133]

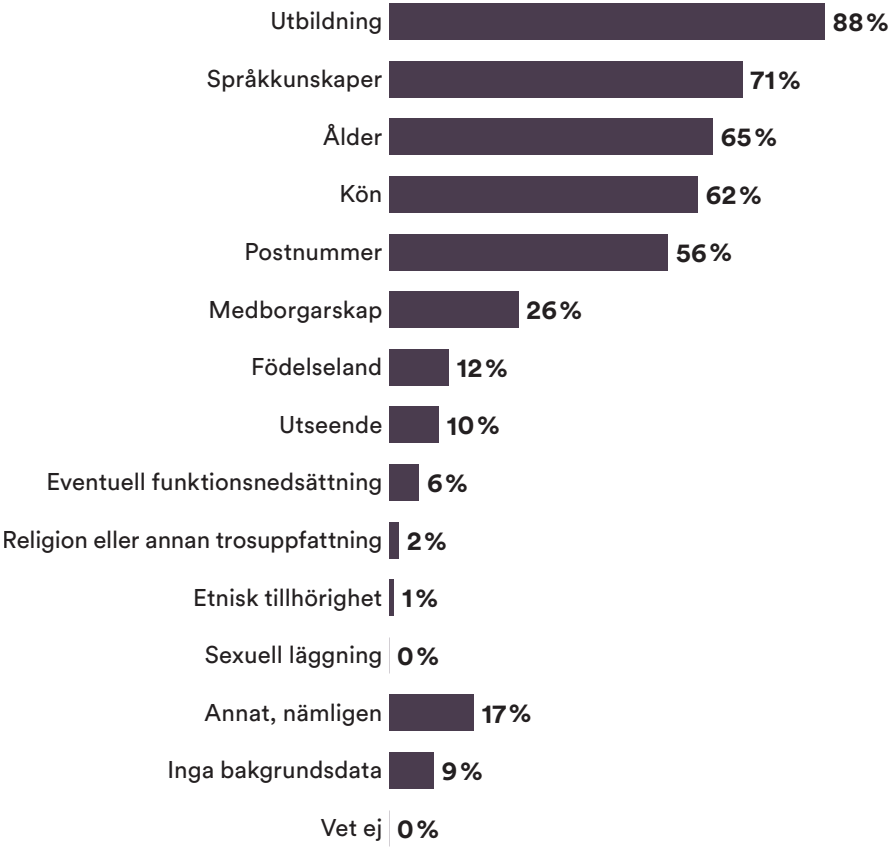
131 För en diskussion om skillnaden på AI-drivet och mänskligt beslutsfattande – varav skala och automation ingår – se Högberg och Larsson (2022).

132 Jämför Koene med flera (2019).

133 Se Crawford med flera (2019). Denna idé finns även uttryckt i Kommissionens förslag på en europeisk AI-förordning, från april 2021 (Artikel 10, punkt 5), där det föreslogs att man i syfte att övervaka, upptäcka och korrigera så kallad bias i relation till högrisk-AI-system, kan processa känsliga personuppgifter.

Figur 5: Fråga 5 till arbetsgivare. Vid en normal rekrytering: vilken av följande bakgrundsdata om era sökande brukar ni ha tillgång till?

Flera svar är möjliga. Antal intervjuer: 100.



När det gäller frågan om vilken bakgrundsdata som arbetsgivarföretagen har vid en normal rekrytering var de vanligaste typerna utbildning (88 procent), språkkunskaper (71 procent), ålder (65 procent), kön (62 procent) och postnummer (56 procent), se figur 5. Respondenterna hade möjligheten att i fritextsvar ange andra kategorier än vad svarsalternativen erbjöd. Av de 17 procent som gjorde så inkom svar som nedan:

> Uppgifterna används av oss för att kunna identifiera och skapa oss en bild av vem personen är.
>
> Uppgiften om medborgarskap används för att kontrollera så att sökande har arbetstillstånd. Ålder och utbildning för att matcha rätt kompetens. Postnumret för att sökande hamnar geografiskt rätt i förhållande till tjänstestället.
>
> Vi vill ha en ökad mix och blandning vad det gäller ålder och kön, vi strävar efter en ökad mångfald i företaget.
>
> Som ett stöd för urval. Det gäller för alla nämnda bakgrundsdata. Inom personlig assistans kan det vara så att vissa vårdtagare vill ha en viss ålder på sin assistent.
>
> Uppgifterna växer samman till en bild av de sökande. För oss är det mer intressant att kompetensen är god hos de sökande än att veta deras sexuella läggning.
>
> Dessa uppgifter används inte alltid men i de fall de används så gör vi detta för att skapa oss en bild av sökandens kompetens.
>
> Vi använder dessa uppgifter för att kunna bedöma den bästa kompetensen för den utlysta tjänsten.

Bristande transparens försvårar möjligheterna att bedöma risken för diskriminering

Bristande transparens är av relevans för risker för diskriminering vid användandet av AI och annat automatiserat beslutsfattande.

Begreppet transparens i sig är ofta otydligt och kan innebära många olika saker. Det kan handla om granskningsbarhet, access (tillgång) till databaser, slutanvändares informationsfärdigheter, eller behovet av dokumentation kring vilka algoritmer som används, vilka mål som styr AI-modellen.[134] Det finns även inneboende intressemotsättningar. Det gäller till exempel behovet av allmän insyn, till exempel för att en tillsynsmyndighet ska kunna granska hur ett AI-system utvecklas, och ett företags intresse av att på en konkurrensutsatt marknad inte avslöja hur deras produkter fungerar.[135] Även den tekniska funktionaliteten kan äventyras av en viss typ av öppenhet som därmed kan missbrukas när det gäller exempelvis system som är menade att flagga oönskade beteenden, som försök till bedrägeri. Forskare argumenterar för att fullständig transparens för AI-system bara bör ges till tillsynsmyndigheter.[136] Den här logiken syns även i EU-kommissionens förslag till AI-förordning, som lägger mycket tonvikt vid granskning och betrodda tredjeparters tillsyn.[137]

Behov av transparens för AI-system i deras utveckling och användning på marknader är också den vanligast utpekade principen i en stor mängd etiska riktlinjer som publicerats under de senaste åren.[138] Det principiella kravet hänger dels ihop med de ovan beskrivna utmaningar med så kallade svarta lådan-modeller, som är svåra eller rentav omöjliga att i detalj förstå hur de har nått sin funktionalitet – och därmed behöver

134 Se till exempel Koene med flera (2019).
135 Larsson och Heintz (2020).
136 De Laat (2018).
137 Mökander med flera (2022).
138 Jobin med flera (2019).

bedömas i relation till risker vid användning, dels rör det sig om ansvarsfördelning. Utan möjligt till ansvarsutkrävande riskerar felaktiga eller diskriminerande AI-system att inte upptäckas och åtgärdas. Mycket av litteraturen kring AI-transparens pekar därmed också på kopplingen till tillförlitlighet. Det vill säga att tilliten till AI- och automatiserade system i någon mån beror på hur transparenta och granskningsbara de är.

När det gäller stora digitala plattformar så är bristande transparens ett uppmärksammat problem för en rad sammanhang, inklusive vilken information som samlas in på webb och av telefoner, i vilket syfte och vad den används till.[139] Det har uppmärksammats att samhällets maktbalanser förskjuts när en rad samhällsfunktioner kommit att kontrolleras av ett fåtal globalt verksamma men samtidigt amerikanskt förankrade teknikkonglomerat.[140] Det inkluderar mycket av arbetslivets digitala praktiker.[141] En följd av detta märks även i både upplägg och resultat av denna undersökning . Resultaten kring sekundäranvändning som har redovisats ovan pekar på ett stort beroende vid rekrytering av tredjeparter som rekryteringsplattformar (till exempel Reachmee och LinkedIn Recruiter) och sociala medier (som Facebook eller LinkedIn, se figur 3) för matchning med riktad annonsering. De flesta större digitala plattformsföretag är väldigt svåra att få tillgång till. Det är därför oklart om jobbmatchning som utförs via dessa tjänster medför eventuella risker för diskriminering.

139 För en kartläggning av så kallade tredjepartskakor på den svenska webben och vad svenskar (inte) förstår och tänker kring det, se Larsson med flera (2021).
140 Van Dijck med flera (2018).
141 Som i tidigare nämnde Lomborgs studie (2022).

Detta tredjepartsberoende kan också medföra en lägre kunskap och medvetenhet om risker för diskriminering vid användning av AI och annat automatiserat beslutsfattande i det svenska arbetslivet.[142] Det orsakar en särskild utmaning vid tillsyn, där även de storskaliga, ofta globala, tekniska plattformarna behöver kunna granskas utifrån (skalbara) risker för diskriminerande utfall.

Otydlig ansvarsfördelning i fråga om diskriminering

Det finns en risk att ansvarsfördelningen i fråga om diskriminering blir otydlig då företag lägger ut sin rekrytering på rekryteringsföretag och där både företag och rekryterare använder en rad tredjepartstjänster för detta ändamål.

På frågan till de större företagen om vem de tror bär ansvar för diskriminering som uppstår till följd av ett företags användning av ett AI-system och annat automatiserat beslutsfattande angav majoriteten, 84 procent, att det är företaget som använder systemet som bär ansvaret. Samtidigt finns här en överhängande risk att varken företagen eller rekryterarna kan kontrollera hur de AI-system som ingår i den digitala matchningen av kandidater via sociala medier eller andra tredjepartstjänster för rekrytering faktiskt säkerställer att de inte diskriminerar. Frågan kompliceras både av AI-systemens inneboende granskningsutmaningar såväl som plattformsföretagens låga transparens.

För de mer avancerade AI-systemen erbjuder ansvarsfördelning särskilda utmaningar, vilket även har påpekats i den så kallade vitbok som EU-kommissionen publicerade 2020 som ett led till förslaget om

142 Vilket argumenteras för i rapporten om Facebooks diskriminerande bostadsförmedling i USA, se Khatry (2020).

AI-förordning.[143] I vitboken anfördes som särskilda utmaningar både föränderligheten hos AI-system, till exempel för produkter som baseras på maskininlärningsberoende programuppdateringar, och ansvarsfördelning på olika platser i en försörjningskedja. Denna problembild syns även i AI-förordningen, där vissa menar att ansvarsfördelningen mellan olika användare inte är tydlig. Det innebär att för arbetsgivare som lägger ut rekrytering och använder externa hjälpmedel för arbetsledning blir det svårt att granska om dessa säkerställer att diskriminering inte sker. Det blir en utmaning vem och vilken del i kedjan som bär ansvaret.

Arbetsgivares kunskap om risker för diskriminering

Detta avsnitt redogör för arbetsgivares kunskap om risker för diskriminering vid tillämpning av AI och annat automatiserat beslutsfattande i samband med rekrytering och arbetsledning.

De huvudsakliga resultaten är:

- Det finns en okunskap på organisatorisk nivå om risker för diskriminering.
- Det finns en okunskap om användning av AI och annat automatiserat beslutsfattande bland rekryteringsföretagen.
- Stora arbetsgivare tycks ha en bristfällig medvetenhet om användningen av AI och annat automatiserat beslutsfattande.

143 Europeiska kommissionen, 19.2.2020, COM (2020) 64 final.

Okunskap på organisatorisk nivå om risker för diskriminering

Respondenterna kommer till övervägande delen från personalavdelningar och förväntas inte fullt ut representera den samlade kunskapen på företagen i fråga om riskerna för diskriminering vid användning av AI och annat automatiserat beslutsfattande. De vet troligen mer om hur dessa system fungerar än den genomsnittliga anställda, men de vet troligtvis mindre om dess tekniska funktionalitet än it-supportpersonalen som installerade eller underhåller dem. Därför representerar de troligen ett tvärsnitt när det gäller kunskapsnivån om dessa system.

Det går att både utforska kunskapsnivåerna och uttryck för okunnighet genom frågornas utformning. Det erbjuds en kompletterande syn inom kunskapssociologin även om det finns en tendens att betrakta okunskap som något dåligt som alltid kan övervinnas med mer lärande. Okunskap kommer helt enkelt i många olika former.[144] Den kan vara aktiv eller passiv, omedveten eller strategisk. Det är möjligt att reflektera över okunskap, men det kommer alltid att finnas möjlighet till händelser så överraskande att de bara kan förstås i efterhand. Okunskap är alltså inte en avvikelse eller undantag, utan något dynamiskt närvarande vid allt mänskligt beslutsfattande. En sådan syn på okunnighet manar till att tänka mer dynamiskt kring kunskap och medvetenhet om risker med AI och annat automatiserat beslutsfattande, hur de uppstår och vad de kan innebära.

144 Gross (2007).

Rekryteringsföretagen och arbetsgivarföretagen tillfrågades om de aktivt förebygger diskriminering i sin användning av AI eller andra former av automatiserat beslutsfattande. Detta är den fråga som tydligast kunde peka på organisatorisk kunskap om diskrimineringsrisker med AI och annat automatiserat beslutsfattande.[145]

I intervjuerna med rekryterare anger 3 av 8 att de har interna riktlinjer, 2 av 8 att de kvalitetssäkrar systemen, medan 3 av 8 anger att de inte använder AI och annat automatiserat beslutsfattande.

Inget av de 8 rekryteringsföretag som svarade på frågan angav något exempel på en extern tredje part (till exempel en konsult specialiserad mot diskrimineringsfrågor) som de har använt sig av, mer än internutbildningar för att undvika fördomar.

Av de sex rekryteringsföretag som svarade på den öppna frågan om vilket stöd eller kunskapsbehov de har för att förebygga risker för diskriminering i AI-system eller system för automatiserat beslutsfattande eller automatiserat beslutsstöd svarade en att de inte vet, fyra att det inte är aktuellt och en att de använder det i så liten skala att inga risker för diskriminering föreligger (ej i tabell). De fyra respondenter som har angett högre frekvens än i mycket låg utsträckning eller inte alls på fråga 1 ovan fick också svara på frågan om de tycker att man bör informera en kandidat om att man använder ett AI-system och annat automatiserat beslutsfattande för att sålla fram rätt kandidat. Tre av dessa svarade att de inte vet medan en svarade jakande, men att de inte gör det i dag (ej i tabell).

145 Organisatorisk kunskap är ett flitigt använt koncept som har införlivats i många företags kvalitetsledningssystem genom så kallad ISO 9001-certifiering, Se Wilson och Campbell (2020).

Av de 100 arbetsgivarföretag som svarade på enkäten fick 95 frågan om de som arbetsgivare aktivt förebygger diskriminering i sin användning av AI eller andra former av automatiserat beslutsfattande. Ett filter användes för att säkerställa att denna fråga endast ställdes till de som rapporterade att de använde AI och annat automatiserat beslutsfattande, eller som rapporterade att de använde en teknik som uppfyllde vår definition av AI och annat automatiserat beslutsfattande. Detta innebar att även de som inte var medvetna om sin användning av AI och annat automatiserat beslutsfattande tillfrågades om vilka processer deras företag har på plats för att minska diskrimineringsrisker.

Omkring hälften, 46 procent, av de tillfrågade på denna fråga svarade att de inte använder AI och annat automatiserat beslutsfattande, se figur 6. Detta är många fler än vad de svarade på den direkta frågan om AI-användning, men ändå avsevärt färre än de som angav att de använder en tjänst som uppfyller vår definition av AI. Den stora majoriteten av de återstående respondenterna bekräftade att de använder en eller flera policyer eller metoder som syftar till att minska risken för diskriminering; endast 2 procent svarade att de inte gör någonting eftersom de förväntar sig att leverantören ska hantera dessa risker, 2 procent svarade att de inte har tänkt på dessa risker och 1 procent att de inte vet.

Vid första anblicken indikerar dessa resultat att organisatorisk kunskap om diskrimineringsrisken med AI och annat automatiserat beslutsfattande är god, särskilt när det finns en medvetenhet om att dessa tekniker används. Totalt 47 procent svarade att de har interna riktlinjer och rutiner för att förebygga diskriminerande effekter, 35 procent att diskriminerande effekter förebyggs genom kvalitetssäkringsprocesser och 32 procent att de har bedömt att deras system är i enlighet med diskrimineringslagen, se figur 6.

Figur 6: Fråga 7 till arbetsgivare. Förebygger ni aktivt diskriminering i er användning av AI eller andra former av automatiserat beslutsfattande som arbetsgivare?

Flera svar är möjliga. Antal intervjuer: 95.

Totalt 53 av arbetsgivarna svarade på fritextfrågan om vilket stöd/kunskapsbehov de har för att förebygga risker för diskriminering i AI-system och annat automatiserat beslutsfattande. Behoven var blandade, och i följande urval av citat kan man utläsa det vanligen uttryckta att inget behov finns, men även att de anser sig ha tillfredsställande juridisk och teknisk expertis, utbildar ledningsgrupp och medarbetare, men även betonandet av företagets värderingar som garant.

> Vi har inget behov/ej aktuellt, då vi inte har något AI-system.
>
> Det viktiga för oss är att uppgifterna inte strider mot de värderingar vi har som företag.
>
> Vi skulle gärna ha mer utbildning bland personal i ledande positioner och i ledningsgruppen gällande diskrimineringsfrågor just för att minimera risken för diskriminering i företaget.
>
> Vår juridiska avdelning samt HR-avdelning och IT-avdelning har tillsammans koll så att reglerna gällande diskriminering efterlevs och inte överskrids av oss.
>
> Vi är en stor koncern och har en hel stab för att säkerställa och förebygga. Staben inkluderar GDPR-ansvarig, IT-chef och jurister. Vi har ganska bra stöd.
>
> Vi har kontrollerat våra system i samråd med vår juridiska avdelning för att förebygga risker för diskriminering.
>
> Vi känner att vi har bra koll kring frågorna om diskriminering.

En stor majoritet (85 procent) angav att de inte har upphandlat ett AI-system eller liknande. Fritextsvar på frågan om hur de har säkerställt att deras upphandlade system inte är diskriminerande pekar dels mot interna processer, dels mot tillit till leverantörens hantering av systemen. Svaren pekar även mot en föreställning om att frånvaron av explicit data om diskrimineringsgrunder gör att verktygen inte kan diskriminera:

> Att vi gått igenom alla frågor i förväg innan mötet med berörd person.
>
> Via internrevisioner från olika leverantörer.
>
> Via upphandlingsprocessen där vi jämfört system och haft detta kriterium i åtanke.
>
> Vi litar på att leverantören har god kunskap i ämnet och har diskrimineringssäkrat sin programvara.
>
> Vi lagrar inte informationen. Erfarenhet och kompetens gäller endast. Systemet plockar bort riskerna för diskriminering.
>
> Via egna internkontroller samt i upphandlingsprocesserna med leverantören av systemet.
>
> Verktygen gör inga urval utifrån de kriterierna utan enbart utifrån erfarenhet och kompetens.

Det framgår dock inte av resultaten hur väl lämpade dessa förebyggande åtgärder är för de särskilda riskerna för diskriminering som finns med AI och annat automatiserat beslutsfattande. Riktlinjer och rutiner är en mjuk form av styrning, som kan syfta till att främja god praxis eller sunda företagsvärderingar utan att exakt specificera hur en rättvis rekrytering och personalledning ska upprätthållas. På liknande sätt används kvalitetssäkringsmekanismer extremt brett av företag för att förbättra anpassningen mellan deras dokumenterade och faktiska metoder. Den dominerande standarden för ledningssystem, ISO 9001, låter företag fastställa sin egen definition av kvalitet och har inga krav på

icke-diskriminering.[146] Det är rimligt att anta – särskilt med tanke på den låga individuella medvetenheten om risker för diskriminering vid användandet av AI och annat automatiserat beslutsfattande – att alla dessa tre åtgärder är generella i den meningen att de inte är inriktade på de särskilda risker som AI och annat automatiserat beslutsfattande utgör. Därmed inte sagt att de kommer att vara helt ineffektiva för att minska dessa risker, bara att de troligen inte är tillräckliga.[147]

Majoriteten av de stora arbetsgivarna tar själva ansvar för hanteringen av diskriminerande risker. Totalt 2 procent av de tillfrågade medgav att de lämnat dessa risker i händerna på sina leverantörer, och bara 12 procent angav att de hade anlitat en tredje part för att granska deras användning av automationsverktyg. Det var en stark majoritet som svarade företaget som använder systemet (84 procent) när de tillfrågades direkt vem de anser bär ansvaret för diskrimineringsrisken med AI och annat automatiserat beslutsfattande. Även om det är viktigt att stora arbetsgivare har åtgärder för att bedöma och minimera riskerna med deras arbetsredskap finns det flera anledningar till att de, när det gäller AI, inte kan göra detta på egen hand. Dessa tekniker kräver inte bara en hög nivå av specialiserad kunskap för att granskas, det finns också gränser för vad man rimligen kan veta om dem. Detta är delvis en fråga om deras komplexitet, men också på grund av den omfattande användningen av tredjepartssystem och tjänster som slutanvändaren sällan har full tillgång till. Företag kan hävda att de är ansvariga för resultatet av dessa verktyg, men den tekniska, organisatoriska och till och med juridiska bördan kan vara svår att reda ut.

146 Ponte med flera (2011).

147 Forskningslitteraturen efterlyser också mer specifika metoder för att mildra ”etiska risker”, se Hunkenschroer och Luetge (2022). För en sådan utvärderingsmetod om AI-risker, etik och transparens, se Felländer med flera (2022).

Okunskap om användning av AI och annat automatiserat beslutsfattande bland rekryteringsföretagen

Vanliga användare av dessa tekniker verkar sällan vara medvetna om att det de använder faktiskt kan klassificeras som AI-stödda system och annat automatiserat beslutsfattande. Vi har tidigare argumenterat för att när AI och annat automatiserat beslutsfattande blir vanligare blir deras användning vardaglig och mindre medvetandegjord som AI. Men det finns också en annan förklaring till detta fenomen. Det sätt som de flesta människor använder AI överensstämmer inte med hur dessa tekniker representeras i populärkulturen och rapporteras om av media. Det som förstås som AI verkar förbli i framkant av den tekniska utvecklingen. AI är bara den allra senaste tekniken (till exempel i skrivande stund, ChatGPT), eller något som är bortom räckhåll som avbildas i science fiction-filmer. Det är något som för alltid finns kvar i den nära framtiden.[148]

En majoritet av rekryteringsföretagen, 6 av 8, tror också att de kommer att skaffa ett AI-baserat system eller annat system för automatiserat beslutstöd för rekrytering inom de kommande tre åren. Så, i korthet, AI uppfattas som en nära framtidsfråga, medan medvetenheten om hur de system de redan använder är automatiserade och i viss mån AI-understödda tycks vara låg.

Låg medvetenhet om användningen av AI och annat automatiserat beslutsfattande är dock inte samma sak som låg medvetenhet om risker för diskriminering i fråga om dessa teknologier. Det är möjligt att trots att de inte tänker på dem som AI och annat automatiserat beslutsfattande, så kan deras användare ändå förstå dessa verktyg som potentiellt partiska och kapabla att producera ojämlika och diskriminerande resultat.

148 Dourish och Bell (2011). Detta har också diskuterats i termer av en ”AI-effekt” eller den ”udda paradoxen”, se Stone med flera (2016). I europeiska regleringssammahang med anledning av både etiska riktlinjer och den föreslagna AI-förordningen diskuteras detta i Larsson (2021).

Det har länge varit känt att när tekniska verktyg, standarder och system används kommer man inte bara att bli beroende av dem utan man kommer också sätta sin tillit till dem. Tekniska system är inte mänskliga och verkar ofta vara befriade från mänskliga fördomar och agerar alltid på ett konsekvent och korrekt sätt, och ges därför auktoritet.[149] Om användarna av AI och annat automatiserat beslutsfattande har blivit så bekväma med verktygen att de inte känner igen dem som AI och annat automatiserat beslutsfattande, då är det troligt att de också har utvecklat ett implicit förtroende för den information och de beslutsstöd som dessa system ger. Detta bådar dock inte gott för en varaktig medvetenhet om deras risker för diskriminering.

Ytterligare bevis för detta påstående tillhandahålls av de typer av risker som människor tenderar att associera med dessa tekniker. Tidigare studier har noterat att oro för existentiella risker (till exempel rädslor för en skenande intelligens eller robotapokalypsen) uttrycks mycket oftare i den offentliga sfären än de som rör vardagliga risker (till exempel intrång i privatlivet eller så kallade fake news). Den allmänna medvetenheten om de diskriminerande riskerna med AI och annat automatiserat beslutsfattande kan därmed antas vara låg. Även om man kan hysa förhoppningar om att en utbildning inom hr-professionerna höjer medvetenheten om risker för diskriminering på arbetsplatsen än genomsnittet inom en befolkning, betyder det inte nödvändigtvis att medvetenheten inkluderar vardagligare applikationer av AI och annat automatiserat beslutsfattande.

149 Porter (2005).

Bristande medvetenhet om användningen av AI och annat automatiserat beslutsfattande hos stora arbetsgivare

Av de 71 företag som rapporterade att de inte alls, eller i mycket låg utsträckning, använder dessa tekniker i sina rekryteringsmetoder (fråga 1, figur 1), använder 52 en digital rekryteringsplattform, 38 använder sociala medier för att rekrytera och 15 använder automatisk cv-filtrering. Av de 76 företag som rapporterade att de inte använder dem för att hantera personal (fråga 2, figur 2), använder 13 automatisk skiftschemaläggning och 9 automatisk sortering och rangordning av personal för att hjälpa till att fastställa löner eller fatta anställningsbeslut.

Det är möjligt att några av de tillfrågade skulle ha ändrat sitt svar på dessa första frågor när de hade svarat på alla andra frågor i enkäten, om de hade fått chansen. Vissa av dessa arbetsgivare anger att de har åtgärder på plats för att hantera deras användning av verktygen. Totalt 26 av de 71 företag som rapporterade utebliven användning av AI och annat automatiserat beslutsfattande vid rekrytering svarade på detta sätt (och 31 av de 76 företag som rapporterade icke-användning för personalledning).

Det är inte betryggande med bristen på medvetenhet om användningen av AI och annat automatiserat beslutsfattande hos stora arbetsgivare. Å ena sidan förväntas många av dem som hävdar icke-användning ändå ha kvalitetskontroller och mekanismer som kan hålla de mest allvarliga riskerna för diskriminering i schack. Å andra sidan tyder det på fortsatt låg medvetenhet av användning, särskilt inom företag med mellan 150–249 och 250–499 anställda.

Sammanfattande resultat

Undersökningen syftade till att studera i vilken utsträckning och i vilka sammanhang som arbetsgivare i Sverige använder sig av vissa AI och annat automatiserat beslutsfattande i arbetslivet, och identifiera vilka risker för diskriminering som finns inbyggda i de tekniska lösningarna. Syftet var även att undersöka hur medvetna arbetsgivare är om dessa risker och vad de gör för att minimera dem. I det avslutande avsnittet lyfts undersökningens sammanfattande resultat fram.

Skillnad mellan upplevd och faktisk användning

Resultaten från undersökningen visar att medvetenheten om den vardagliga användningen av AI och annat automatiserat beslutsfattande är låg. Undersökningen visar en tydlig tudelning i resultaten hos både rekryteringsföretag och större arbetsgivarföretag. Det gäller medvetenhet om den primära användningen av AI och annat automatiserat beslutsfattande kontra den sekundära användningen av AI och annat automatiserat beslutsfattande som ingår i externutvecklade tjänster som företagen använder.

En majoritet svarade att det var i låg utsträckning eller inte alls på en direkt fråga om i vilken utsträckning de system de använder för att hantera rekrytering inkluderar olika former av automatiserat beslutsfattande eller automatiserat beslutsstöd. Detsamma gäller för den direkta frågan om de använder system med AI-funktionalitet. Frekvensen ökar dock väsentligt om till exempel rekryteringsföretagen tillfrågas om de använder AI-system eller andra verktyg för att matcha kandidater med riktad annonsering mot deras profil. Eftersom detta är en tydlig referens till de sociala medier som enligt en rad studier är vanliga i jobbmatchningssammanhang, tyder det på att respondenterna trots allt använder sig av ett slags AI-funktionalitet med stort inslag av automation för att hantera rekrytering. Detta även om de inte förefaller uppmärksamma det som AI och annat automatiserat beslutsfattande.

En majoritet anger också att de får in uppgifter om lämpliga kandidater från sociala medieprofiler.

Liknande resultat visar sig i den högre frekvens som anges av de större företagen när de tillfrågas om specifika verktyg och tjänster som används i rekryteringsprocesser, såsom rekryteringsplattformar, sociala medier för matchning med mera. Denna dynamik innebär att det är analytiskt relevant att skilja på primär och sekundär användning av AI och annat automatiserat beslutsfattande.

Det är också tydligt att arbetsgivarna i undersökningen inte är medvetna om att de redan nu använder sig av AI och annat automatiserat beslutsfattande, till exempel i det löpande arbetsledande personalarbetet såsom vid hanteringen av lönesystem med mera.

Brist på transparens och möjlighet till granskning leder till risker för diskriminering

Det blir i förlängningen viktigt med transparens och möjlighet till granskning samt medvetenhet hos användare både om vad det är för tjänster de använder och vilka inneboende risker de kan ha, oavsett grad eller typ av automatiserat beslutsfattande i digitala system – baserat på mer komplex AI-metodik eller enklare automation.

Resultaten från undersökningen visar att stora plattformar och globala teknikbolag utgör ett slags automatiserad digital infrastruktur för både rekrytering och arbetsledning, dock med en bristande transparens som är relevant för att kunna identifiera risker för diskriminering. Här finns en komplexitet mellan stora företag med rekryteringsbehov som delegerar delar till rekryteringsföretag samtidigt som båda parter använder sig av datadrivna digitala tredjepartstjänster som automatiserar matchning med inslag av AI-understödda prediktionsmodeller.

Generella förebyggande åtgärder är otillräckliga

Undersökningen visar att 47 procent av arbetsgivarna visserligen uppger att de har interna riktlinjer och rutiner när det gäller det förebyggande arbetet med att förhindra diskriminering vid användandet av AI och annat automatiserat beslutsfattande. Därutöver anger 35 procent att de kvalitetssäkrar sina system för att förebygga diskriminerande effekter eller utvärderar dessa för att de ska överensstämma med diskrimineringslagen. Samtidigt anger de att de inte använder AI-stödda system och annat automatiserat beslutsfattande för rekrytering eller arbetsledning. En rimlig tolkning är att förebyggandet ofta förstås i en generell mening, inte explicit kopplad till AI och annat automatiserat beslutsfattande.

Sammantaget pekar undersökningen på att stora arbetsgivare ser sig själva som ansvariga för hanteringen av risker för diskriminering. Det var en tydlig majoritet som svarade företaget som använder systemet när de tillfrågades direkt om vem de anser bär ansvaret för risker för diskriminering med AI och annat automatiserat beslutsfattande. Det är dock oklart hur väl lämpade dessa förmodat mer generella åtgärder är för de diskrimineringsrisker som följer av just användning av AI och annat automatiserat beslutsfattande. Det stora användandet av tredjepartssystem som samtidigt inte medvetet sågs som automatiserade beslutsstöd eller som inkluderande AI-funktionalitet talar för att vissa risker förenade med denna användning kan falla mellan stolarna och inte uppmärksammas eftersom kunskapen är låg.

Mer forskning om risker för diskriminering vid användning av AI och annat automatiserat beslutsfattande behövs

Det finns behov av fördjupande studier på flera områden när det gäller kunskap kring risker för diskriminering vid användning av AI och annat automatiserat beslutsfattande. Överlag finns det forskningsluckor i hela fältet kring AI och annat automatiserat beslutsfattande i relation till risker för diskriminering. Det gäller särskilt i relation till hur tillväxten ser ut när det gäller befintliga praktiker, såväl som tekniska frågor om att hantera bias, och rättsliga tolkningar i ljuset av AI-fältets utveckling.

Den digitaliserade och datainsamlande arbetsledningen och dess eventuella risker för diskriminering väcker fler frågor som det inte finns helt tillfredsställande svar på utifrån denna undersökning. Mer studier behövs för att avgöra vilka typer av lösningar svenska företag använder inom detta område, och i vilken utsträckning och hur de använder AI och annat automatiserat beslutsfattande i det.

Även om det sannolikt kommer att variera i vilken utsträckning digitala plattformtjänster använder automatisering och AI-baserade lösningar som beslutsstöd, har marknadsledaren inom detta område, LinkedIn, publicerat viss information om hur de gör detta.[150] Det finns ett behov av mer kunskap om de digitala plattformarnas roll inom svensk rekryteringspraxis och deras hantering av risker för diskriminering i relation till AI och annat automatiserat beslutsfattande. Mer uppmärksamhet kan alltså riktas mot den roll de storskaliga digitala plattformarna spelar för diskrimineringsfrågorna i allmänhet, och vilka åtgärder de vidtar för att minska riskerna för diskriminering i synnerhet.

150 Se LinkedIn (den 9 oktober 2018), Forbes (den 2 juli 2020).

Undersökningen vittnar om att det samtidigt finns förväntningar hos många arbetsgivare på en nära förestående tillväxt i AI-understödd rekrytering. Det gör att detta fält behöver mer fördjupning och bevakning framöver i relation till risker för diskriminering, sammantaget med att undersökningen är ett nedslag i ett snabbrörligt fält som beror både på teknikutveckling av AI-metoder med omvittnade utmaningar i fråga om transparens och att europeisk reglering samtidigt är att vänta. Både den primära och sekundära användningen förväntas öka, med till exempel integration av stora språkmodeller i vardagliga jobbverktyg som ordbehandlare och sökmotorer. Det betyder att användningen sannolikt kommer att uppfattas som ännu mer vardaglig och därför delvis passera obemärkt – även om riskerna för diskriminering samtidigt kan kvarstå.

Kapitel 6 Sammanfattande resultat och DO:s slutsatser

Denna rapport är en slutredovisning av det uppdrag som DO har fått av regeringen gällande risker för diskriminering i arbetslivet vid användningen av AI och annat automatiserat beslutsfattande.

Som vi har beskrivit i inledningen är syftet med uppdraget att kartlägga vilka risker för diskriminering som användandet av AI och annat automatiserat beslutsfattande kan medföra och i vilken utsträckning och i vilka sammanhang som arbetsgivare använder sådana tekniska lösningar.[151] Vidare bör kunskapsnivån belysas hos arbetsgivare om förhållanden som kan leda till diskriminering när det gäller AI och annat automatiserat beslutsfattande.

I arbetet med regeringsuppdraget har DO anlitat två forskargrupper. Den ena forskargruppen har genomfört en systematisk forskningsöversikt för att undersöka vilka risker för diskriminering som har identifierats i forskning vid användning av AI och annat automatiserat beslutsfattande. Den andra forskargruppen fick i uppgift att undersöka i vilken utsträckning och i sammanhang som arbetsgivare använder sig av AI och annat automatiserat beslutsfattande och vilken kunskapsnivå de har om risker för diskriminering. DO vill förtydliga att de forskargrupper som DO har anlitat har delvis använt en annan definition av diskriminering än diskrimineringslagens definition.

151 Arbetsmarknadsdepartementet (2022), Regeringsbeslut 2022-06-07, A2022/00877.

Med utgångspunkt i dessa två undersökningar kan DO övergripande konstatera att AI och annat automatiserat beslutsfattande används i ökande omfattning av arbetsgivare i dag, även om dessa inte alltid är medvetna om vilken teknik de använder. Likaså kan vi konstatera att AI och annat automatiserat beslutsfattande medför risker för diskriminering och att arbetsgivare i dag endast till viss del är medvetna om dessa risker. Det innebär att kunskapsnivån behöver höjas både om när man som arbetsgivare använder AI och annat automatiserat beslutsfattande och hur risker för diskriminering kan förebyggas.

I detta avslutande kapitel sammanfattar DO forskarnas resultat och redovisar våra slutsatser i fråga om åtgärder för att minska risken för diskriminering vid användningen av AI och annat automatiserat beslutsfattande.

Motstridiga forskningsresultat och forskningsluckor

Forskningsöversikten visar att det finns stora forskningsluckor om risker för diskriminering och AI och annat automatiserat beslutsfattande i arbetslivet. Det framkommer att det vetenskapliga underlaget är tämligen begränsat.

Forskningen pekar i många fall åt olika håll. Vissa artiklar visar på möjligheter med AI och annat automatiserat beslutsfattande när det gäller att förebygga diskriminering. Samtidigt lyfter andra artiklar fram riskerna för diskriminering när AI och annat automatiserat beslutsfattande används inom samma område och på liknande sätt.

Flera studier visar att system baserade på AI och annat automatiserat beslutsfattande kan minska risken för diskriminering i samband med urval av kandidater i rekryteringsprocesser. Det finns forskning som pekar på att urvalet kan bli mer neutralt gällande variabler som kön, ålder och etnisk tillhörighet när det genomförs med hjälp av system baserade på AI

och annat automatiserat beslutsfattande. Detta gäller i synnerhet om systemen aktivt har utformats för att säkerställa att urvalet är neutralt gällande dessa variabler. Samtidigt visar forskning på det motsatta: att användning av AI och annat automatiserat beslutsfattande i samband med annonsering och urval också riskerar att leda till diskriminering.

Återspegling av ojämlikheter på arbetsmarknaden

Flera forskningsartiklar lyfter fram att risker för diskriminering kan uppkomma eftersom den data som systemen tränas på återspeglar historiska och befintliga ojämlikheter och grupprelaterade skillnader i arbetslivet. Systemen riskerar då att fatta eller föreslå beslut som reproducerar dessa gruppskillnader. Denna grundläggande omständighet riskerar exempelvis att leda till att systemen tolkar denna data som att etablerade mönster av över- och underrepresentation mellan grupper inom olika branscher eller på olika befattningsnivåer återspeglar skillnader i kompetens och lämplighet.

Automation bias – att okritiskt acceptera utfallet av systemen

Av forskningsöversikten framkommer även att mänskliga beslutsfattare som agerar i relation till system baserade på AI och annat automatiserat beslutsfattande riskerar att bli mindre noggranna och kritiska. I stället tenderar de att uppvisa så kallad automation bias. Det innebär att de okritiskt accepterar utfallet av systemen som något neutralt, objektivt och värderingsfritt. Risken är att diskriminerande utfall som systemen utvecklar inte upptäcks och korrigeras om användarna inte är medvetna om att dessa system – precis som alla andra metoder för informationsbearbetning och beslutsfattande – har såväl för- som nackdelar och måste granskas kritiskt.

Forskning pekar också på att system baserade på AI och annat automatiserat beslutsfattande kan användas för att osynliggöra eller legitimera diskriminering. Arbetsgivare kan motivera och rättfärdiga utfall och beslut även om diskriminering har förekommit genom att hänvisa till utfallen från systemen som icke-diskriminerande och sakliga.

AI-systemen kan leda till systematisk diskriminering

Vidare visar forskningsöversikten att system baserade på AI och annat automatiserat beslutsfattande kan åstadkomma diskriminerande utfall konsekvent och systematiskt. Det kommer att finnas en stor variation i hur diskriminering på grund av mänskliga beslutsfattare ser ut och vilka konsekvenser denna diskriminering får i enskilda fall, även om systemen skulle vara mindre diskriminerande än mänskliga beslutsfattare i de enskilda fallen. Sådan diskriminering har en föränderlighet som drivs av en rad omständigheter, vilket har med såväl person- som situationsfaktorer att göra. Den diskriminering som sker riskerar att bli mer enhetlig, konsekvent och systematisk om ett begränsat antal system baserade på AI och annat automatiserat beslutsfattande, eller flera olika system som fungerar på liknande sätt, dominerar inom ett område. Risken utgörs av att en ökad användning av sådana system skulle göra diskriminerande processer och utfall än mer systematiska och konsekventa.

Skillnad mellan upplevd och faktisk användning av AI och annat automatiserat beslutsfattande

Undersökningen av både utbredning och kunskapsnivåer är avhängig hur AI och annat automatiserat beslutsfattande förstås. Begreppen AI och annat automatiserat beslutsfattande är svårfångade och i ständig rörelse. Dess varierande innebörd kan också ha betydelse för hur arbetsgivare upplever sin användning av AI och annat automatiserat beslutsfattande,

det vill säga vad de lägger i begreppen och vad de uppfattar som AI och annat automatiserat beslutsfattande.

Undersökningen visar att få arbetsgivare är medvetna om i vilken utsträckning de använder sig av AI och annat automatiserat beslutsfattande. Det finns en åtskillnad mellan upplevd och faktisk användning av AI och annat automatiserat beslutsfattande hos svenska arbetsgivare.

Medvetenhet om den primära användningen av AI och annat automatiserat beslutsfattande, det vill säga lösningar som har utvecklats i betydande utsträckning av de företag som använder dem, uppgavs vara låg hos både rekryteringsföretag och större arbetsgivarföretag. Samtidigt uppgavs den sekundära användningen av AI och annat automatiserat beslutsfattande som ingår i externt utvecklade tjänster som företagen använder vara hög. Det framkommer liknande resultat när de större företagen tillfrågas om specifika verktyg och tjänster som används i rekryteringsprocesser (rekryteringsplattformar, sociala medier för matchning med mera).

Det blir också tydligt i undersökningen att arbetsgivare inte är medvetna om att de redan nu använder sig av AI och annat automatiserat beslutsfattande, till exempel i det löpande arbetsledande personalarbetet såsom vid hanteringen av lönesystem med mera. Det innebär att det finns en risk att dessa tekniska lösningar både är väl integrerade i vardagen och ändå ibland okritiskt och omedvetet betrodda av de personer som använder dem. Det kan betyda att risker för diskriminering inte uppmärksammas eller hanteras, och att ansvar hänskjuts till någon annan.

Brist på transparens och möjlighet till granskning leder till risker för diskriminering

Undersökningen visar att stora företag med rekryteringsbehov upphandlar tjänster från rekryteringsföretag som i sin tur använder sig av tredjeparts-system, det vill säga stora plattformar som ägs av globala teknikbolag. Detta kan försvåra transparens och möjlighet till granskning då de datadrivna digitala tredjepartstjänster som automatiserar matchning är svåra att granska. Bristande transparens kan i sin tur leda till risker för diskriminering.

Undersökningen pekar på att systemen inte kan undkomma mänskliga fördomar och ojämlikheter, eftersom systemen utvecklas och även tränas av människor och på data som representerar mänskligt språk och samhälle. Det innebär att AI och annat automatiserat beslutsfattande alltid medför risker för diskriminering. AI- systemen behöver därmed granskas och hanteras för att nå de kostnadseffektiva och precisa matchningseffekter som förväntas uppnås vid rekrytering via dessa tredjepartslösningar.

Otillräcklig kunskap om risker för diskriminering

Det finns behov av fördjupande studier på flera områden när det gäller kunskap kring risker för diskriminering vid användning av AI och annat automatiserat beslutsfattande. Överlag finns det forskningsluckor i hela fältet kring AI och annat automatiserat beslutsfattande i relation till risker för diskriminering. Det gäller särskilt i relation till hur tillväxten ser ut när det gäller befintliga praktiker, såväl som tekniska frågor som att hantera bias.

Det finns behov av mer kunskap om de digitala plattformarnas roll inom svensk rekryteringspraxis och deras hantering av risker för diskriminering i relation till AI och annat automatiserat beslutsfattande, även om det sannolikt kommer att variera i vilken utsträckning digitala plattforms-tjänster använder automatisering och AI-baserade lösningar som beslutsstöd.

DO:s slutsatser

AI-systemen måste bli mer transparenta för att motverka diskriminering

AI-systemen måste bli mer transparenta då bristande transparens gör det svårt att identifiera risker för diskriminering. En tillsynsmyndighet ska kunna granska hur ett AI-system utvecklas. Samtidigt är de flesta större digitala plattformsföretag väldigt svåra att få tillgång till. Det orsakar en särskild utmaning vid tillsyn, där även de storskaliga, ofta globala, tekniska plattformarna behöver kunna granskas utifrån (skalbara) risker för diskriminerande utfall.

Dessutom måste ansvarsfördelningen bli tydligare kring vem som bär ansvar för risker för diskriminering. Arbetsgivaren har ansvar för att ingen utsätts för diskriminering. Men det finns en överhängande risk att varken företagen eller rekryterarna kan kontrollera hur de AI-system som ingår i den digitala matchningen av kandidater via sociala medier eller andra tredjepartstjänster för rekrytering faktiskt säkerställer att de inte diskriminerar. Utan möjlighet till ansvarsutkrävande riskerar felaktiga eller diskriminerande AI-system att inte upptäckas och åtgärdas.

Det behövs också rättsliga tolkningar i ljuset av AI-fältets utveckling som skulle kunna belysa risker för diskriminering. Arbetet inom EU med att ta fram en förordning om AI är ett steg i riktning mot att reglera vissa AI-system. För att betona att skyddet mot diskriminering gäller parallellt med AI-förordningen har DO påtalat att det i AI-förordningen bör framgå att ett godkännande enligt förordningen inte påverkar leverantörers och användares ansvar för diskriminering som sker genom sådana AI-system.[152]

152 Diskrimineringsombudsmannen, 2021-06-23.

Risker för systematisk diskriminering och representationsbias i AI-system måste förhindras

Den ökade användningen av AI-system kan bidra till diskriminerande processer och utfall som blir mer systematiska och konsekventa och som drabbar många individer. Risker för systematisk diskriminering måste därför förhindras.

AI-system bygger också på data som återspeglar historiska och befintliga ojämlikheter och grupprelaterade skillnader i arbetslivet, vilket kan leda till strukturella fördomar. De har även inbyggda representationsbias, vilket innebär att träningsdata inte representerar alla grupper. Det påverkar i sin tur AI-systemens precision negativt för dessa grupper. Det finns också ojämlika och diskriminerande strukturer i samhället som reproduceras i exempelvis språk och texter som används som träningsdata. Eftersom AI-modeller tränas på stora datamängder kan man inte utesluta att alla fördomar de har också kommer att vara utbredda. Detta ökar risken för diskriminering i stor omfattning, vilket också kan vara svårt att upptäcka.

För AI-utvecklare och AI-användare, arbetsgivare, blir det därmed viktigt att upptäcka, åtgärda och förhindra sådan så kallad bias i AI-systemen. Det kan till exempel ske genom att säkerställa att fler kompetenser eller underrepresenterade grupper är inkluderade i designprocesser och dataurval. Det kan i sin tur säkerställa att eventuella diskriminerande effekter i systemens tillämpning motverkas.

Arbetsgivare måste öka sin medvetenhet om AI och annat automatiserat beslutsfattande och arbeta med aktiva åtgärder

Att man som arbetsgivare känner till att man använder AI och annat automatiserat beslutsfattande är avgörande för att arbeta förebyggande för att motverka diskriminering vid tillämpning av tekniken.

Resultaten i rapporten visar att den faktiska användningen bland arbetsgivare är högre än den upplevda användningen av sådana tekniska lösningar. Det tycks vara svårt för arbetsgivare att vara helt säkra på när de använder system som bygger på någon form av AI eller automatiserat beslutsfattande. Kunskapen om vad AI och annat automatiserat beslutsfattande är, och medvetenheten om när arbetsgivare använder tekniken, måste därmed öka. I detta sammanhang är det även viktigt att arbetsgivare aktivt arbetar för att motverka automation bias. Det vill säga att tilltron till de automatiserade besluten leder till ett passivt, okritiskt och oreflekterat förhållningssätt hos den beslutande arbetsgivaren.

Kunskapsnivån hos arbetsgivare är inte tillfredsställande när det gäller de risker för diskriminering som finns vid användningen av AI och annat automatiserat beslutsfattande. En slutsats är därför att arbetsgivare måste ta ett ökat ansvar för att motverka diskriminering vid användningen av AI och annat automatiserat beslutsfattande och följa diskrimineringslagen.

Arbetsgivare har en skyldighet att leva upp till diskrimineringslagens krav på aktiva åtgärder, oavsett vilka tekniska lösningar de använder. Det innebär att arbetsgivare ska bedriva ett förebyggande och främjande arbete för att motverka diskriminering och på annat sätt verka för allas lika rättigheter och möjligheter oavsett diskrimineringsgrund inom området arbetsliv.

Arbetet ska ske kontinuerligt i fyra steg. Arbetsgivare ska undersöka, analysera, åtgärda och följa upp samt utvärdera risker för diskriminering inom området arbetsliv gällande

- arbetsförhållanden
- bestämmelser om praxis om löner och andra anställningsvillkor
- rekrytering och befordran
- utbildning och övrig kompetensutveckling
- möjlighet att förena förvärvsarbete med föräldraskap.

Avslutningsvis kan vi konstatera att utvecklingen av AI och annat automatiserat beslutsfattande i dag och för lång tid framöver kommer att vara ett område som DO och andra likabehandlingsorgan i Europa behöver följa. Tillämpningen av AI och annat automatiserat beslutsfattande kan över tid både innebära minskade liksom ökade risker för diskriminering. Hur AI-systemens diskriminerande effekter fungerar i praktiken, inte minst i en svensk kontext, behöver därmed undersökas vidare. Även mer kvalitativ forskning behöver bedrivas för att avgöra vilka typer av lösningar svenska företag använder inom detta område, liksom i vilken utsträckning och hur de använder AI och annat automatiserat beslutsfattande och vilka risker för diskriminering detta kan medföra.

Referenser

Kapitel 1

AI HLEG (2019b). En definition av AI – viktigaste förmågor och discipliner.

AI HLEG (2019c). Etiska riktlinjer för tillförlitlig AI.

Arbetsmarknadsdepartementet (2022). Regeringsbeslut 2022-06-07, A2022/00877.

Diskrimineringsombudsmannen (2022). Transparens, Träning och Data. Myndigheters användning av AI och automatiserat beslutsfattande samt kunskap om risker för diskriminering. Rapport 2022:1.

Drage, Eleanor och Mackereth, Kerry (2022). Does AI Debias Recruitment? Race, Gender, and AI's Eradication of Difference. Philosophy and Technology, 35(4), sida 857–936.

Equinet (2022) Dutch Childcare Allowance Scandal: the importance of Investigation Powers. Bryssel

Equinet (2020a). Meeting the New Challenges to Equality and Non-discrimination from Increased Digitisation and the use of Artificial Intelligence. Executive Summary. Bryssel.

Equinet (2020b). Regulating for an Equal AI: A New Role for Equality bodies. Meeting the new challenges to equality and non-discrimination from increased digitisation and the use of Artificial Intelligence. A Good Practice Guide. Bryssel.

Equinet (2020c). A Good Practice Guide. Bryssel.

Europarådet (2019). Unboxing Artificial Intelligence: 10 steps to protect Human Rights.

Europeiska kommissionen. Meddelande från kommissionen till Europaparlamentet, Europeiska rådet, rådet, Ekonomiska och sociala kommittén och Regionkommittén om artificiell intelligens för Europa, 25.4.2018, COM (2018) 237 final.

Europeiska unionens byrå för grundläggande rättigheter, FRA, (2020). Getting the future right – Artificial intelligence and fundamental rights (europa.eu). Wien.

Insight Intelligence (2022). Svenska folket och AI. Svenska folkets attityder till artificiell intelligens 2022. Stockholm. Genomfört i samarbete med Bolagsverket, Lunds universitet, Postnord och Göteborgs universitet.

Larsson, Stefan och Heintz, Fredrik (2020). Transparency in Artificial Intelligence. Internet Policy Review, 9(2).

Larsson, Stefan och Ledendal, Jonas (2022). AI i offentlig sektor: Från etiska riktlinjer till lagstiftning. I: de Vries, Katja och Dahlberg, Mattias (redaktörer De Lege årsbok 2021: Law, AI & Digitalization). Iustus förlag.

Medium (den 20 juni 2022). "Guidelines for journalists and editors about reporting on robots, AI, and computers" (medium.com) av Ben Shneiderman [senast besökt den 2 juni 2023].

Medium (den 18 april 2022). "On NYT Magazine on AI: Resist the Urge to be Impressed" (medium.com), av Emily M. Bender [senast besökt den 2 juni 2023].

Myndigheten för digital förvaltning, DIGG, (2020). Främja den offentliga förvaltningens förmåga att använda AI. Delrapport i regeringsuppdraget I201.

Regeringen (2018). Den nationella inriktningen för artificiell intelligens. N2018.14.

Riksrevisionen (2020:22). Automatiserat beslutsfattande i statsförvaltningen – effektivt, men kontroll och uppföljning brister. Stockholm.

Roehl, Ulrik (2022). Understanding Automated Decision-Making in the Public Sector: A Classification of Automated, Administrative Decision-Making. I: Juell-Skielse, Gustav, Lindgren, Ida och Åkesson, Maria (redaktörer). Service Automation in the Public sector. Concepts, Empirical Examples and Challenges. Cham: Springer Nature Switzerland.

Russell, Stuart och Peter Norvig (2022). Artificial Intelligence. A modern Approach, 4th ed. Storbritannien: Pearson Education Limited.

Shestakova, Veronika. (2021). Best Practices to Mitigate Bias and Discrimination in Artificial Intelligence. Performance Improvement, 60(6), sida 6–11.

Slota, Stephen C., Fleischmann, Kenneth R., Greenberg, Sherri, Verma, Nitin, Cummings, Brenna, Li, Lan och Shenefiel, Chris (2020). Good systems, bad data? Interpretations of AI hype and failures. Proceedings of the Association for Information Science and Technology, 57(1), e275.

Kapitel 2

Equinet (2023) AI Act ensuring Equality in Europe: Our Recommendations – Equinet (equineteurope.org), hämtad 2023-07-27.

Europakommissionen (2023). A European approach to artificial intelligence | Shaping Europe's digital future (europa.eu), hämtad 2023-07-27.

Europaparlamentet Nyheter (2023). AI-regler: Vad vill Europaparlamentet lagstifta om? (europarl.europa.eu), hämtad 2023-07-27.

Europaparlamentet (2023). Legislative Train Schedule Artificial intelligence act. In "A Europe Fit for the Digital Age" (europarl.europa.eu), hämtad 2023-07-27.

Europarådet (2023). Council of Europe and Artificial Intelligence (coe.int), hämtad 2023-07-27

Europeiska unionens officiella tidning (2012). Europeiska unionens stadga om de grundläggande rättigheterna (eur-lex.europa.eu), sida 391-407, hämtad 2023-07-27.

Kapitel 3

Diskrimineringsombudsmannen (2022). Transparens, Träning och Data. Myndigheters användning av AI och automatiserat beslutsfattande samt kunskap om risker för diskriminering. Rapport 2022:1.

Proposition 2007/08:95. Ett starkare skydd mot diskriminering.

Proposition 2002/03:65. Ett utvidgat skydd mot diskriminering.

Diskrimineringslagen (2008:567).

Kapitel 4

Agrawal, Ajay, Gans, Joshua och Goldfarb, Avi (2018). Prediction machines: The simple economics of artificial intelligence. Harvard Business Review Press.

Ajunwa, Ifeoma (2018). Algorithms at Work: Productivity Monitoring Applications and Wearable Technology as the New Data-Centric Research Agenda for Employment and Labor Law. St. Louis University Law Journal, 63(1), sida 21–53.

Ali, Muhammed, Sapiezynski, Piotr, Bogen, Miranda, Korolova, Aleksandra, Mislove, Alan och Rieke, Aaron (2019). Discrimination through Optimization : How Facebook's Ad Delivery Can Lead to Biased Outcomes. Proceedings of the ACM on Human-Computer Interaction, 3, artikelnummer 199, sida 1–30.

Angrave, David, Charlwood, Andy, Kirkpatrick, Ian, Lawrence, Mark och Stuart, Mark (2016). HR and analytics: why HR is set to fail the big data challenge. Human Resource Management Journal, 26(1), sida 1-11.

Arbetsmarknadsdepartementet (2022), Regeringsbeslut 2022-06-07, A2022/00877.

Bales, Richard, A och Stone, Katherine V. W. (2020). The Invisible Web at Work- Artificial Intelligence and Electronic Surveillance in the Workplace. Berkeley Journal of Employment & Labor Law, 41(1), sida 1–62.

Benjamin, Ruha (2019). Race After Technology: Abolitionist Tools for the New Jim Code. Cambridge and Medford: Polity Press.

Birhane, Abeba (2021). Algorithmic injustice: A relational ethics approach. Patterns, 2(2), sida 1–9.

Björklund, Fredrik, Bäckström, Martin och Wolgast, Sima (2012). Company norms affect which traits are preferred in job candidates and may cause employment discrimination. The Journal of Psychology, 146(6), sida 579–594.

Bogen, Miranda (2019). All the Ways Hiring Algorithms Can Introduce Bias. Harvard Business Review Digital Articles, 2–5.

Bolander, Thomas (2019). What do we loose when machines take the decisions? Journal of Management & Governance, 23(4), sida 849–867.

Bucher, Eliane, Fieseler, Christiane och Lutz, Christoph (2019). Mattering in digital labor. Journal of Managerial Psychology, 34(4), sida 307–324.

Busuioc, Madalina (2021). Accountable Artificial Intelligence: Holding Algorithms to Account. (2021*)*. Public Administration Review, 81(5), sida 825–836.

Charlwood, Andy och Guenole, Nigel (2022). Can HR adapt to the paradoxes of artificial intelligence? Human Resource Management Journal, 32(4), sida 729–742.

Cheng, Maggie M. och Hackett, Rick D. (2021). A critical review of algorithms in HRM: Definition, theory, and practice. Human Resource Management Review,31(1), sida 1–14.

Chungyalpa, W. och Karishma, T. (2016). Best practices and emerging trends in recruitment and selection. Journal of Entrepreneurship & Organization Management, 5(2), sida 1-5.

Daugherty Paul. R. och Wilson H. James (2018). Human+ machine: reimagining work in the age of AI. Boston; Harvard Business Press.

Drage, Eleanor och Mackereth, Kerry (2022). Does AI Debias Recruitment? Race, Gender, and AI's Eradication of Difference. Philosophy and Technology, 35(4), sida 857–936.

Duckworth, Angela Lee, Quinn, Patrick. D., Lynam, Donald. R., Loeber, Rolf, Stouthamer-Loeber, Magda och Smith, Edward. E. (2011). Role of test motivation in intelligence testing. Proceedings of the National Academy of Sciences of the United States of America, 108(19), sida 7716–7720.

Fountain, Jane. E. (2022). The moon, the ghetto and artificial intelligence: Reducing systemic racism in computational algorithms. Government Information Quarterly, 39(2).

Goretzko, David och Finja Israel, Laura Sophia (2022). Pitfalls of machine learning-based personnel selection: Fairness, transparency, and data quality. Journal of Personnel Psychology, 21(1), sida 37–47.

Griesbach, K., Reich, A., Elliott-Negri, L. och Milkman, R. (2019). Algorithmic control in platform food delivery work. Socius, 5, 1–15.

Haesevoets, Tessa, De Cremer, David, Dierckx, Kim och Van Hiel, Alain (2021). Human-machine collaboration in managerial decision making [Article]. Computers in Human Behavior, 119.

Hamilton, R. H. och Davison, H. K. (2022). Legal and Ethical Challenges for HR in Machine Learning [Article]. Employee Responsibilities & Rights Journal, 34(1), sida 19–39.

Henrich, Joseph, Heine, Steven J. och Norenzayan, Ara (2010). The weirdest people in the world? Behavioral and Brain Sciences, 33(2–3), sida 61–83.

Hoffmann, Anna Lauren (2019). Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. Information Communication and Society, 22(7), sida 900–915.

Horton, John J. (2017). The Effects of Algorithmic Labor Market Recommendations: Evidence from a Field Experiment. Journal of Labor Economics, 35(2), sida 345–385.

Kappen, Mitchel och Naber, Marnix (2021). Objective and bias-free measures of candidate motivation during job applications. Scientific Reports, 11(1), sida 1–8.

Keith, Melissa. G., Harms, Peter och Tay, Louis (2019). Mechanical Turk and the gig economy: Exploring differences between gig workers. Journal of Managerial Psychology, 34(4), sida 286–306.

Kellogg, Katherine, Valentine, Melissa och Christin, Angèle (2020). Algorithms at work: The new contested terrain of control. Academy of Management Annals, 14(1), sida 366–410.

Kelly-Lyth, Aislinn (2021). Challenging Biased Hiring Algorithms. Oxford Journal of Legal Studies, 41(4), sida 899–928.

Kim, Pauline. T. (2017). Data-Driven Discrimination at Work. William & Mary Law Review, 58(3), sida 857–936.

Kim, Pauline. T. (2020). Manipulating opportunity. Viriginia Law Review, 106(4), sida 867–935.

Kim, Pauline. T. och Bodie, Matthew T. (2021). Artificial Intelligence and the Challenges of Workplace Discrimination and Privacy. ABA Journal of Labor & Employment Law, 35(2), sida 289–315.

Kim, Pauline. T. och Scott, Sharion (2018). Discrimination in Online Employment Recruiting. Saint Louis University Law Journal, 63(1), sida 93–118.

Knocke, Wuokko, Drejhammar, Inga-Britt, Gonäs, Lena och Isaksson, Kerstin (2003). Retorik och praktik i rekryteringsprocessen. I: Arbetslivsinstitutet: Arbetsliv i omvandling.

Koechling, Alina, Riazy, Shirin, Wehner, Maurius Claus och Simbeck, Katharina (2020). Highly Accurate, But Still Discriminatory: A Fairness Evaluation of Algorithmic Video Analysis in the Recruitment Context. Business & Information Systems Engineering.

Kuhn, Max och Johnson, Kjell (2016). Applied predictive modeling. Springer.

Köchling, Alina och Wehner, Maurius Claus (2020). Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. Business Research, 13(3), sida 795–848.

Lambrecht, Anja och Tucker, Catherine (2019). Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. Management Science, 65(7), sida 2966–2981.

Langer, Markus, König, Cornelius. J. och Papathanasiou, Maria (2019). Highly automated job interviews: Acceptance under the influence of stakes. International Journal of Selection & Assessment, 27(3), sida 217–234.

Larose, Daniel T. och Larose, Chantal D. (2014). Discovering knowledge in data: An introduction to data mining (2nd ed.). John Wiley.

Lee, Min Kyung (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. Big Data & Society, 5(1), sida 1–16.

Leicht-Deobald, Ulrich, Busch, Thorsten, Schank, Christoph, Weibel, Antoinette, Schafheitle, Simon, Wildhaber, Isabelle och Kasper, Gabriel (2019). The Challenges of Algorithm-Based HR Decision-Making for Personal Integrity. Journal of Business Ethics, 160(2), sida 377–392.

Mani, Anandi, Mullainathan, Sendhil, Shafir, Eldar och Zhao, Jiaying (2013). Poverty Impedes Cognitive Function. Science, 341(6149), sida 976–980.

Mann, Gideon och O'Neil, Cathy (2016). Hiring Algorithms Are Not Neutral. Harvard Business Review Digital Articles, 2–4.

Martínez, Narao, Vinas, Aranzazu och Matute, Helena (2021). Examining potential gender bias in automated-job alerts in the Spanish market. PLoS ONE, 16(12), sida 1–15.

McDonald, Steve (2011). What's in the old boys network? Accessing social capital in gendered and racialized networks. Social Networks, 33, sida 317–330.

McDonald, Kathleen, Fisher, Sandra och Connelly Catherine (2017). e-HRM systems in support of smart workforce management: an exploratory case study of system success. Electronic HRM in the Smart Era, 87–108.

Mehrabi, Ninareh, Morstatter, Fred, Saxena, Nripsuta, Lerman, Kristina och Galstyan, Aram (2022). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54(6), 1–35.

Melchers, Klaus G. och Basch, Johannes. M. (2022). Fair play? Sex-, age-, and job-related correlates of performance in a computer-based simulation game. International Journal of Selection & Assessment, 30(1), 48–61.

Mitchell, Tom (1997). Machine Learning. McGraw Hill.

Nationalencyklopedin, algoritm. Hämtad 2022-12-06.

Newman, David. T., Fast, Nathanael. J. och Harmon, Derek. J. (2020). When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. Organizational Behavior and Human Decision Processes, 160, sida 149–167.

Office for AI (2019). Understanding artificial intelligence (gov.uk). Retrieved December 5, 2022.

Parent-Rocheleau, Xavier och Parker, Sharon K. (2022). Algorithms as work designers: How algorithmic management influences the design of jobs. Human Resource Management Review, 32(3), sida 1–17.

Peeters, Rik, Giest, Sarah och Grimmelikhuijsen, Stephan (2020). The agency of algorithms: Understanding human-algorithm interaction in administrative decision-making. Information Polity: The International Journal of Government & Democracy in the Information Age, 25(4), sida 507–522.

Quillian, Lincoln, Heath, Anthony, Pager, Devah, Midtbøen, Arnfinn, Fleischmann, Fenella och Hexel, Ole (2019). Do some countries discriminate more than others? Evidence from 97 field experiments of racial discrimination in hiring. Sociological Science, 6, sida 467–496.

Roth, Philip. L., Bobko, Philip, Van Iddekinge, Chad H. och Thatcher, Jason Bennett (2016). Social media in employee-selection-related decisions: a research agenda for uncharted territory. Journal of Management, 42, sida 269–298.

Sajjadiani, Sima, Sojourner, Aaron. J., Kammeyer-Mueller, John. D. och Mykerezi, Elton (2019). Using machine learning to translate applicant work history into predictors of performance and turnover. Journal of Applied Psychology, 104(10), sida 1207–1225.

Savage, David. D. och Bales, Richard (2017). Video Games in Job Interviews: Using Algorithms to Minimize Discrimination and Unconscious Bias. ABA Journal of Labor and Employment Law, 32(2), sida 211–228.

Shestakova, Veronika (2021). Best Practices to Mitigate Bias and Discrimination in Artificial Intelligence. Performance Improvement, 60(6), sida 6–11.

Simbeck Katharina (2019). HR analytics and ethics. IBM Journal of Research and Development, 63 (4/5), sida 1–9.

Skanderson, David M. (2021). Managing Discrimination Risk of Machine Learning and AI Models. ABA Journal of Labor & Employment Law, 35(2), sida 339–360.

Suen, Hung-Yue., Chen, Mavis-Yi Ching och Lu, Shih-Hao (2019). Does the use of synchrony and artificial intelligence in video interviews affect interview ratings and applicant attitudes? Computers in Human Behavior, 98, sida 93–101.

Tambe, Prasanna, Cappelli, Peter och Yakubovich, Valery (2019). Artificial intelligence in human resources management: Challenges and a path forward. California Management Review, 61 (4), sida 15–42.

Tursunbayeva, Aizhan, Pagliari, Claudia, Di Lauro, Stefano och Antonelli, Gilda (2022). The ethics of people analytics: risks, opportunities and recommendations. Personnel Review, 51(3), sida 900–921.

Vandenberghe, Christian (1999). Organizational Culture, Person-Culture Fit, and Turnover: A Replication in the Health Care Industry. Journal of Organizational Behavior, 20(2), sida 175–184.

van Esch, Patrick, Black, J. Stewart och Ferolie, Joseph (2019). Marketing AI recruitment: The next phase in job application and selection. Computers in Human Behavior, 90, sida 215–222.

Varma, Arup, Dawkins, Cedric och Chaudhuri, Kaushik (2022). Artificial intelligence and people management: A critical assessment through the ethical lens. Human Resource Management Review.

Williams, Betsy Anne, Brooks, Catherine. F. och Shmargad, Yotam (2018). How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications. Journal of Information Policy, 8, sida 78–115.

Wood, Alex. J., Graham, Mark, Lehdonvirta, Vili och Hjorth, Iris (2019). Good gig, bad gig: autonomy and algorithmic control in the global gig economy. Work, Employment and Society, 33(1), sida 56–75.

Woods, Stephen Anthony, Ahmed, Sara, Nikolaou, Ioannis, Costa, Ana Christina och Anderson, Neil Robert (2020). Personnel selection in the digital age: a review of validity and applicant reactions, and future research challenges. European Journal of Work and Organizational Psychology. 29 (1), sida 64–77.

Yang, Jenny R. (2021). Adapting Our Anti-Discrimination Laws to Protect Workers' Rights in the Age of Algorithmic Employment Assessments and Evolving Workplace Technology. ABA Journal of Labor & Employment Law, 35(2), sida 207–240.

Yarger, Lynette, Cobb Payton, Fay och Neupane, Bikalpa (2019). Algorithmic equity in the hiring of underrepresented IT job candidates. Online Information Review, 44(2), sida 383–395.

Kapitel 5

Böcker, antologier, kapitel och artiklar

van Bekkum, Marvin och Zuiderveen Borgesius, Frederik (2023). Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception? Computer Law & Security Review, 48, 105770.

Benjamin, Ruha (2019). Race After Technology: Abolitionist Tools for the New Jim Code. Cambridge and Medford: Polity Press.

Birhane, Abeba (2022). The unseen Black faces of AI algorithms. Nature, 610(7932), sida 451–452.

Bolukbasi, Tolga, Chang, Kai-Wei, Zou, James Y., Saligrama, Venkatesh och Kalai, Adam T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. I: Daniel D. Lee, Ulrike von Luxburg, Roman Garnett, Masashi Sugiyama och Isabelle Guyon (redaktörer), NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems New York: Curran Associates Inc. sida 4349–4357.

Buolamwini, Joy och Gebru, Timnit (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency sida 77–91.

Caliskan, Aylin, Bryson, Joanna och Narayanan, Arvind (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), sida 183–186.

Crawford, Kate (2021). Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. New Haven: Yale University Press.

Crawford, Kate, Dobbe, Roel, Dryer, Theodora, Fried, Genevieve, Green, Ben, Kaziunas, Elizabeth, Kak, Amba, Mathur, Varoon, McElroy, Erin, Nill Sánchez, Andrea, Raji, Deborah, Lisi Rankin, Joy, Richardson, Rashida, Schultz, Jason, Myers West, Sarah, och Whittaker, Meredith (2019). AI Now 2019 Report. New York City: AI Now Institute.

Crawford, Kate och Paglen, Trevor (2021). Excavating AI: The politics of images in machine learning training sets. AI & Society, 36(4), sida 1105–1116.

Delfanti, Alessandro (2021). The Warehouse: Workers and Robots at Amazon. London: Pluto Press.

De Laat, Paul B. (2018). Algorithmic decision-making based on machine learning from big data: can transparency restore accountability? Philosophy & technology, 31(4), sida 525–541.

D'Ignazio, Catherine, och Klein, Lauren F. (2020). Data Feminism. Cambridge and London: The MIT Press.

Dourish, Paul, och Bell, Genevieve (2011). Divining a Digital Future: Mess and Mythology in Ubiquitous Computing. Cambridge and London: The MIT Press.

Ford, Martin (2015). Rise of the Robots: Technology and the threat of a jobless future. New York: Basic Books.

Felländer, Anna, Rebane, Jonathan, Larsson, Stefan, Wiggberg, Mattias och Heintz, Fredrik (2022). Achieving a Data-driven Risk Assessment Methodology for Ethical AI. Digital Society, 1(2), 13.

Gasser, Urs och Almeida, Virgilio. A. (2017). A layered model for AI governance. IEEE Internet Computing, 21(6), sida 58–62.

Gaucher, Danielle och Justin Friesen, and Aaron C. Kay (2011). Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality. Journal of Personality and Social Psychology, July 2011, Vol 101(1), sida 109–28.

Gillespie, Tarleton (2018). Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.

Gross, Matthias (2007). The Unknown in Process: Dynamic Connections of Ignorance, Non-Knowledge and Related Concepts. Current Sociology, 55(5), sida 742–759.

Haider, Jutta och Sundin, Olof (2019). Algoritmernas roll i plattformssamhället: Vad är algoritmer, och vad gör dem till så viktiga komponenter i plattformssamhället? I: Andersson Schwarz, Jonas och Larsson, Stefan (redaktörer). Plattformssamhället: Den digitala utvecklingens politik, innovation och reglering. Stockholm: Fores (sida 90–113).

Hunkenschroer, Anna Lena och Luetge, Christoph (2022). Ethics of AI-enabled recruiting and selection: A review and research agenda. Journal of Business Ethics, 178(4), sida 977–1007.

Högberg, Charlotte och Larsson, Stefan (2022). AI and patients' rights: Transparency and information flows as situated principles in public health care. De Lege, 2021, sida 401–429.

Insight Intelligence (2022). Svenska folket och AI. Svenska folkets attityder till artificiell intelligens 2022. Stockholm. Genomfört i samarbete med Bolagsverket, Lunds universitet, Postnord och Göteborgs universitet.

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

Jones, Phil (2021). Work without the worker: labour in the age of platform capitalism. London: Verso.

Kaun, Anne (2022). Suing the algorithm: the mundanization of automated decision-making in public services through litigation. Information, Communication & Society, 25(14), 2046–2062.

Kaun, Anne, Lomborg, Stine, Pentzold, Christian, Allhutter, Doris, och Sztandar-Sztanderska, Karolina (2023). Crosscurrents: Welfare. Media, Culture & Society.

Khatry, Sarah (2020). Facebook and Pandora's box: How using big data and artificial intelligence in advertising resulted in housing discrimination. Applied Marketing Analytics, 6(1), sida 37–45.

Koene, Ansgar, Clifton, Chris, Hatada, Yohko, Webb, Helena och Richardson, Rashida (2019). A governance framework for algorithmic accountability and transparency (Study No. PE 624.262) Panel for the Future of Science and Technology, Scientific Foresight Unit (STOA), European Parliamentary Research Service.

Koivisto, Ida (2022). The Transparency Paradox. Oxford University Press.

Larsson, Stefan (2019). Sju nyanser av transparens: Om artificiell intelligens och ansvaret för digitala plattformars samhällspåverkan. I: Andersson Schwarz, Jonas och Larsson, Stefan (redaktörer). Plattformssamhället: Den digitala utvecklingens politik, innovation och reglering. Stockholm: Fores, sida 277–313.

Larsson, Stefan (2019). The Socio-Legal Relevance of Artificial Intelligence. Droit et Société, 103(3), sida 573–593.

Larsson, Stefan (2020). On the governance of artificial intelligence through ethics guidelines. Asian Journal of Law and Society, 7(3), sida 437–451.

Larsson, Stefan (2021). AI in the EU: Ethical Guidelines as a Governance Tool i Bakardjieva Engelbrekt, Antonia, Leijon, Karin, Michalski, Anna och Oxelheim, Lars (redaktörer). The European Union and the Technology Shift. Cham, Switzerland: Palgrave Macmillan.

Larsson, Stefan, Anneroth, Mikael, Felländer, Anna, Felländer-Tsai, Li, Heintz, Fredrik, Ångström, Rebecka, och Åström, Fredrik (2019). HÅLLBAR AI: Inventering av kunskapsläget för etiska, sociala och rättsliga utmaningar med artificiell intelligens. Stockholm: AI Sustainability Center.

Larsson, Stefan, Haresamudram, Kashyap, Högberg, Charlotte, Lao, Yucong, Nyström, Alex, Söderlund, Kasia och Heintz, Fredrik (2023). Four Facets of AI Transparency, i Lindgren, Simon, (redaktörer) Handbook of Critical Studies in Artificial Intelligence, Edward Elgar Publishing.

Larsson, Stefan och Heintz, Fredrik (2020). Transparency in Artificial Intelligence. Internet Policy Review, 9(2).

Larsson, Stefan, Jensen-Urstad, Anders och Heintz, Fredrik (2021). Notified but Unaware: Third-Party Tracking Online. Critical Analysis of Law, 8(1), sida 101–120.

Leavy, Susan (2018). Gender bias in artificial intelligence: the need for diversity and gender theory in machine learning. In Proceedings of the 1st International Workshop on Gender Equality in Software Engineering (GE '18). Association for Computing Machinery, New York, NY, USA, 14–16.

Levy, Karen (2023). Data Driven: Truckers, Technology, and the New Workplace Surveillance. Princeton and Oxford: Princeton University Press.

Lomborg, Stine (2022). Everyday AI at Work: Self-tracking and automated communication for smart work. In Everyday Automation Routledge, sida 126–139.

Miller, Tim (2019). Explanation in Artificial Intelligence: Insights from the social sciences. Artificial intelligence, 267, sida 1–38.

Morondo Taramundi, D. (2022). Discrimination by Machine-Based Decisions: Inputs and Limits of Anti-Discrimination Law. I: Custers, B & Fosch-Villaronga, E (redaktörer). Law and Artificial Intelligence: Regulating AI and Applying AI in Legal Practice The Hague: TMC Asser Press. sida 73–85.

Moss, Haley (2020). Screened out onscreen: Disability discrimination, hiring bias, and artificial intelligence. Denv. L. Rev., 98, 775.

Myers West, Susan, Whittaker, Meredith och Crawford, Kate (2019). Discriminating Systems: Gender, Race, and Power in AI. New York: AI Now Institute.

Mökander, Jacob, Axente, Maria, Casolari, Federico och Floridi, Luciano (2022). Conformity assessments and post-market monitoring: A guide to the role of auditing in the proposed European AI regulation. Minds and Machines, 32(2), sida 241–268.

Noble, Safiya Umoja (2018). Algorithms of Oppression: How Search Engines Reinforce Racism. New York: New York University Press.

Neri, Hugo och Cozman, Fabio (2020). The role of experts in the public perception of risk of artificial intelligence. AI & Society, 35(3), sida 663–673.

van Nuenen, Tom, Ferrer, Xavier, Such, Jose M. och Coté, Mark (2020). Transparency for whom? assessing discriminatory artificial intelligence. Computer,53(11), sida 36–44.

Ponte, Stefano, Gibbon, Peter och Vestergaard, Jakob (redaktörer) (2011). Governing through Standards: Origins, Drivers and Limitations. Basingstoke and New York: Palgrave Macmillan.

Porter, Tony (2005). Private Authority, Technical Authority, and the Globalization of Accounting Standards. Business and Politics, 7(3), sida 427–440.

Roehl, Ulrik (2022). Understanding Automated Decision-Making in the Public Sector: A Classification of Automated, Administrative Decision-Making. In Juell-Skielse, Gustav, Lindgren, Ida och Åkesson, Maria (redaktörer). Service Automation in the Public sector. Concepts, Empirical Examples and Challenges. Cham: Springer Nature Switzerland.

Rosenblat, Alex (2018). Uberland: How Algorithms Are Rewriting the Rules of Work. Oakland, CA: University of California Press.

Russell, Stuart (2019). Human Compatible: AI and the Problem of Control. London: Penguin Books.

Russell, Stuart och Peter Norvig (2022). Artificial Intelligence. A modern Approach, 4th ed. Storbritannien: Pearson Education Limited.

Pasquale, Frank (2015). The black box society: The secret algorithms that control money and information. Harvard University Press.

Shankar, Shreya, Halpern, Yoni, Breck, Erik, Atwood, James, Wilson, Jimbo och Sculley, D. (2017). No classification without representation: Assessing geodiversity issues in open data sets for the developing world.

Stone, Peter med flera (2016). Artificial Intelligence and Life in 2030. Stanford University: Study Panel.

Trewin, Shari, Basson, Sara, Muller, Michael, Branham, Stacy, Treviranus, Jutta, Gruen, Daniel och Manser, Erich (2019). Considerations for AI fairness for people with disabilities. AI Matters, 5(3), sida 40–63.

van den Broek, Elmira, Sergeeva, Anastasia, Huysman Vrije, Marleen (2021). When the Machine Meets the Expert: An Ethnography of Developing AI for Hiring. MIS Quarterly, 45(3), sida 1557–1580.

van Dijck, José, Poell, T. och De Waal, M. (2018). The Platform Society: Public values in a connective world. Oxford University Press.

Wernberg, J. (2019). I den svarta lådan: Plattformsekonomier och digitalise-ring. I: Jonas Andersson Schwarz och Stefan Larsson (redaktörer Plattformssamhället: Den digitala utvecklingens politik, innovation och reglering Stockholm: Fores. sida 22–60.

Whittaker, Meredith, Alper, Meryl, Bennett, Cynthia L, Hendren, Sara, Kaziunas, Liz, Mills, Mara, Ringel Morris, Meredith, Rankin, Joy, Rogers, Emily, Salas, Marcel, och Myers West, Sarah (2019). Disability, bias, and AI. New York: AI Now Institute.

White, James M. och Lidskog, Rolf (2022). Ignorance and the regulation of artificial intelligence. Journal of Risk Research, 25(4), 488–500.

Wilson, John P. och Campbell, Larry (2020). ISO 9001:2015: the evolution and convergence of quality management and knowledge management for competitive advantage. Total Quality Management & Business Excellence, 31(7–8), sida 761–776.

Woodcock, Jamie och Graham, Mark (2020). The Gig Economy: A Critical Introduction. Cambridge and Medford: Polity.

Xenidis, Raphaële och Senden, Linda (2020). EU non-discrimination law in the era of artificial intelligence: Mapping the challenges of algorithmic discrimination. I: Ulf Bernitz med andra (redaktörer). General Principles of EU law and the EU Digital Order. Kluwer Law International, 2020, sida 151–182.

Offentligt tryck och lagtexter

Arbetsmarknadsdepartementet (2022). Regeringsbeslut 2022-06-07, A2022/00877.

Diskrimineringslag (2008:567).

Europaparlamentets och rådets förordning (EU) 2022/2065 av den 19 oktober 2022 om en inre marknad för digitala tjänster och om ändring av direktiv 2000/31/EG (förordningen om digitala tjänster).

Europeiska kommissionen. Förslag till Europarlamentets och rådets förordning om harmoniserade regler för artificiell intelligens (rättsakt om artificiell intelligens) och om ändring av vissa unionslagstiftningsakter, 21.4.2021, COM (2021) 206 final.

Europeiska kommissionen, 19.2.2020, COM(2020) 64 final.

Europeiska kommissionen. Meddelande från kommissionen till Europaparlamentet, Europeiska rådet, rådet, Ekonomiska och sociala kommittén och Regionkommittén om artificiell intelligens för Europa, 25.4.2018, COM (2018) 237 final.

Regeringen (2018). Den nationella inriktningen för artificiell intelligens. N2018.14.

Myndighetsrapporter

AI HLEG (2019a). Ethics Guidelines for Trustworthy AI. Bryssel: EU-kommissionen.

AI HLEG (2019b). En definition av AI – viktigaste förmågor och discipliner.

AI HLEG (2019c). Etiska riktlinjer för tillförlitlig AI.

Department of Justice (2022). Justice Department Secures Groundbreaking Settlement Agreement with Meta Platforms, Formerly Known as Facebook, to Resolve Allegations of Discriminatory Advertising (justice.gov) [senast besökt den 12 juni 2023].

Diskrimineringsombudsmannen (2022:1). Transparens, Träning och Data. Myndigheters användning av AI och automatiserat beslutsfattande samt kunskap om risker för diskriminering.

DIGG (2020). Främja den offentliga förvaltningens förmåga att använda AI. Delrapport i regeringsuppdraget I201.

Riksrevisionen (2020:22). Automatiserat beslutsfattande i statsförvaltningen – effektivt, men kontroll och uppföljning brister. Stockholm.

SCB (den 18 januari 2023). ”Antalet anställda i snabbväxande storföretag kommer att öka” (scb.se) [senast besökt den 5 juni 2023].

Nyhetssändningar och texter hämtade från internet

Forbes (den 2 juli 2020) How AI Is Impacting Operations at LinkedIn av Kathleen Walch (forbes.com) [senast besökt den 5 juni 2023].

LinkedIn (den 9 oktober 2018). An Introduction to AI at LinkedIn (engineering.linkedin.com) [senast besökt den 5 juni 2023].

Kapitel 6

Arbetsmarknadsdepartementet (2022). Regeringsbeslut 2022-06-07, A2022/00877.

Diskrimineringsombudsmannen, 2021-06-23. Yttrande över Europeiska kommissionens förslag till förordning om harmoniserade regler för artificiell intelligens, Proposal for a Regulation laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts (COM(2021)206), LED 2021/290, handling 3.

Bilaga 1: Metodbeskrivning forskningsöversikt

Som tidigare har beskrivits har syftet med den aktuella forskningsöversikten varit att besvara följande frågeställningar:

1. Hur ser det aktuella forskningsläget ut gällande riskerna för diskriminering vid användning av artificiell intelligens och annat automatiserat beslutsfattande i arbetslivet?
2. Hur ser det aktuella forskningsläget ut gällande användning av artificiell intelligens och annat automatiserat beslutsfattande som kan förebygga diskriminering i arbetslivet?
3. Vilka är kunskapsluckorna inom forskningsfältet?

Målsättningen med forskningsöversikten har därför varit att systematiskt sammanställa resultaten från all tillgänglig forskning som belyser frågeställningarna ovan givet de avgränsningar som specificeras i forskningsöversiktens syften. För att göra det har en så kallad systematisk forskningsöversikt genomförts. En systematisk forskningsöversikt innebär att sammanställa alla forskningsresultat som finns tillgängliga om en viss fråga vid en viss tidpunkt och bygger på att man söker och bearbetar vetenskaplig litteratur på ett sätt som möjliggör att forskningsöversiktens syften uppnås. Den procedur och de avvägningar som har tillämpats i den aktuella översikten presenteras nedan.

Avgränsningar, sökord och söksträngar

I ett första steg måste sökningens avgränsningar, sökord och söksträngar definieras. Med avgränsningar avses i detta sammanhang vilken sorts vetenskaplig litteratur som ska inkluderas och med sökord menas vilka begrepp man ska söka på i aktuella databaser för att fånga upp samtliga relevanta artiklar på området. Söksträngar syftar på hur dessa sökord ska sättas samman i sökningar som blir så träffsäkra som möjligt i relation till forskningsöversiktens syften.

Samtliga beslut och överväganden i relation till söktermer och avgränsningar har diskuterats och beslutats i dialog mellan forskarna och representanter för Diskrimineringsombudsmannen, och alla steg har dokumenterats.

Först måste alltså de avgränsningar som kan härledas ur forskningsöversiktens syfte definieras och tydliggöras. I den aktuella forskningsöversikten har följande avgränsningar gjorts:

- Resultat från forskning har definierats som resultat publicerade i vetenskapliga tidskrifter med kollegial granskning (så kallad peer-review).
- Aktuell har definierats som att artiklarna ska vara publicerad under de senaste fem åren, det vill säga 2017–2022.

Nästa steg är som sagt att omsätta forskningsöversiktens nyckelbegrepp till sökningar i relevanta databaser över vetenskaplig litteratur på ett sätt som säkerställer att forskningsöversiktens syften kan uppnås. Databaserna i den aktuella översikten valdes ut för att uppnå god täckning gällande relevanta forskningsfält och tidskrifter. För att skapa bra sökningar behöver de så kallade sökorden, det vill säga de ord som används i sökningen, vara utformade på ett sätt så att man fångar upp alla relevanta artiklar på området, trots att begreppsanvändning och beteckningar ofta kan variera. I den aktuella forskningsöversikten var

det därför avgörande att skapa träffsäkra sökord i relation till diskriminering, risker för diskriminering och förebyggande åtgärder mot diskriminering, AI och annat automatiserat beslutsfattande samt arbetslivet. För att identifiera dessa sökord har forskarna tagit hjälp av informationsspecialister vid Lunds universitetsbibliotek, jämfört med andra internationella översikter på liknande teman samt undersökt vilka nyckelord och begrepp som används i relevant vetenskaplig litteratur på området.

Sökorden i forskningsöversikten organiseras därefter i så kallade sökblock som syftar till att ringa in översiktens huvudbegrepp. I den aktuella forskningsöversikten definierades följande sökblock:

1. diskriminering (risker och förebyggande)
2. artificiell intelligens och annat automatiserat beslutsfattande
3. arbetslivet.

Sökblock 1 formulerades för att fånga upp både artiklar som handlar om diskriminering i vid bemärkelse och artiklar som i stället formulerar sin ansats positivt, det vill säga att de snarare använder begrepp som handlar om att motverka diskriminering eller åstadkomma inkludering, jämlikhet och liknande.

Som kan ses ovan valde forskarna att inte skapa sökblock relaterade till de olika diskrimineringsgrunderna. Man införde inte heller att artiklarna explicit skulle handla om risker för missgynnande som kan knytas till någon av diskrimineringsgrunderna i den svenska diskrimineringslagen som ett nödvändigt kriterium för inkludering senare i processen. Detta eftersom översikten omfattar även internationell vetenskaplig litteratur som inte alltid använder samma begrepp och klassifikationer som i den svenska diskrimineringslagen. Dessutom fokuserar många artiklar på det aktuella området inte på diskriminering i relation till specificerade grupper, utan rör i stället

diskriminering och diskriminerande konsekvenser på en mer generell och övergripande nivå.

Det är i detta sammanhang också viktigt att påpeka att forsknings-översikten inte syftar till att identifiera risker och möjligheter i relation till alla former av digitala system eller hjälpmedel som används i arbetslivet. Forskningsöversikten omfattar endast forskning om sådana system som bygger på AI och annat automatiserat beslutsfattande. Forskning som rör exempelvis datoriserade tester eller spel som urvalsinstrument eller användandet av sociala medier för att nå ut till och bedöma potentiella kandidater till tjänster har alltså enbart inkluderats om de undersökta systemen bygger på AI och annat automatiserat beslutsfattande. På liknande sätt har endast studier som rör risker eller möjligheter för diskriminering i arbetslivet i samband med användandet av artificiell intelligens inkluderats. Detta innebär att det som har exkluderats är forskning rörande andra, såväl positiva som negativa, effekter på arbetslivet eller arbetsmarknaden till följd av den ökade användningen av system baserade på AI.

Sökblocken kombinerades i en söksträng (sökblock 1 +2 +3) och sökningar gjordes i databasplattformen Ebscohost, där forskarna samsökte i databaserna Academic Search Complete, Business Source Complete, MEDLINE, APA PsycInfo, och SocINDEX.

Söksträngar och resultat från sökningen redovisas i Tabell 1. Sökningar gjordes i titel, nyckelord och abstract.

Tabell 1. Sökströngar och sökresultat

Sökströngar	Antal träffar i använda databaser
(Discriminat* OR Fair* OR Equal* OR Inclus* OR Equit*) AND (Automat* OR "Automatic decision making OR "Artificial intelligence OR Comput* OR Algorithm* OR Machine OR AI Or ADM OR Robot* OR deep learning OR neural network" OR NLP OR "Natural language processing) AND (Work* OR Employ* OR Hiring OR Recruit* OR Hire OR "human resource management OR HR OR Organization OR Organisation OR Staff OR Personell OR Leadership OR Management OR Promot* OR Job* OR Select* OR advertisement OR "human resource OR applicant OR "people analytics)	Ebscohost (Academic Search Complete, Business Source Complete, MEDLINE, APA PsycInfo, SocINDEX): 4 537 träffar

Inkludering och exkludering

I nästa steg granskade de två forskarna titel och abstract för samtliga träffar i de genomförda sökningarna för att ta ställning till om den aktuella artikeln ska genomgå fulltextgranskning, det vill säga läsas i sin helhet. Titel- och abstractgranskningen syftar till att säkerställa att artikeln uppfyller de uppställda kriterierna för att inkluderas i den aktuella forskningsöversikten.

Titel och abstractgranskningen gjordes avsiktligt något överinkluderande, vilket innebär att artiklar som utifrån titel och abstract var svåra att avgöra om de skulle inkluderas eller inte gick vidare till fulltextläsning för att säkerställa att relevanta studier inte exkluderades.

Titel och abstractgranskning genomfördes av de två ansvariga forskarna oberoende av varandra. Varefter de inkluderade studierna från båda forskarna sammanfördes till en lista och dubbletter rensades ut. Detta resulterade i 89 artiklar som gick vidare till fulltextläsning.

Under fulltextläsningen lästes samtliga artiklar i sin helhet och baserat på denna läsning avgjordes om de kunde inkluderas i det slutliga urvalet av artiklar till forskningsöversikten. Detta urval baserades på samtliga kriterier för inkludering och exkludering som har redovisats tidigare. Även fulltextläsningen genomfördes av de två forskarna oberoende av varandra, varefter de två listor på inkluderade artiklar som detta genererade slogs samman och dubbletter rensades ut. Detta resulterade i ett slutligt urval på 58 artiklar som utgjorde underlaget för forskningsöversikten.

Närläsning och tematisering

I arbetet med forskningsöversikten lästes samtliga artiklar av båda forskarna. Artiklarnas innehåll och resultat diskuterades och tematiserades i relation till forskningsöversiktens syften. För varje artikel granskades också metoden för att även skapa en bild av hur forskningen på området ser ut rent metodologiskt och därigenom bidra till underlaget för de diskussioner som förs i forskningsöversikten om begränsningar i den aktuella litteraturen och behoven av framtida forskning.

Bilaga 2: Metodbeskrivning undersökning av rekryterings-företag och större företag

Stefan Larsson och James White vid Lunds universitet har i samverkan med Novus och i dialog med DO tagit fram en undersökningsdesign. I syfte att säkerställa en god undersökningsdesign genomförde forskarna även sex intervjuer under projektets inledningsfas.

De bidrog till förtydliganden i de två enkäter som riktades mot rekryteringsföretag respektive företag med mer än 100 anställda. Enkäterna togs fram tillsammans med Novus och utvärderades och kommenterades av medarbetare på DO.

För att få kontakt med relevanta rekryteringsföretag utgick uppdragstagarna från en lista från arbetsgivarorganisationen Kompetensföretagen (en del av Almega) över de 50 största bemannings-företagen – som inkluderar rekryteringsföretag – i landet baserat på omsättning. Det beställdes en lista över CTO:s vid dessa rekryterings-företag, det vill säga Chief Technology Officier eller Chief Technical Officer, även kallat it-chef eller teknisk direktör. Dessa kontaktades via telefon för att boka en tid för intervju. Var det så att kontaktuppgifterna till respektive respondent inte var direkt kopplad till personen sökte man kontakt via företagets växel.

En bedömning gjordes att CTO på rekryteringsföretagen har en god insikt i frågor som är centrala för denna undersökning, inte minst med tanke på deras förväntade kunskap om användning av rekryteringsverktyg och databaser. Om CTO var rätt person i denna målgrupp kontrollerades genom att utvärdera efter ett par genomförda intervjuer. Om det vid

kontaktsökandet visade sig att det inte var rätt att intervjua just företagens CTO, så kunde den personen också hänvisa oss till rätt person.

Gällande de större arbetsgivarföretagen bedömdes att ett passande och tillräckligt urval kunde nås genom att beställa fram ett register över företag med mer än 100 anställda, där två separata urval drogs. Urvalen drogs slumpmässigt utan hänsyn till bransch, geografi eller liknande. Urvalsregistret är Statistiska Centralbyråns (SCB) företagsregister. Ena företagsurvalet var med företag med 100–249 anställda, och det andra med 250+ anställda. Givet erfarenheter med tidigare kartläggningar av liknande slag konstaterades att 50 intervjuer var rimligt att sträva efter för varje urvalsgrupp.

Novus genomförde den empiriska datainsamlingen som behövdes för undersökning och riskanalys. Den bestod av två strukturerade undersökningar som genomfördes per telefon.

Målgrupp 1: rekryteringsföretag

Gruppen med rekryteringsföretag bedömdes skulle få en svarsfrekvens om cirka 30 procent med representanter från de 50 största företagen inom sektorn. Detta skulle motsvara cirka 15 intervjuer.

Det tycktes finnas flera skäl till att just denna grupp var svåra att få att ställa upp på intervju, trots återkommande försök. Telefonväxeln hänvisade vid ett flertal tillfällen till epost som första kontakt, men även i vissa fall att företaget har en policy om att inte svara på undersökningar. En del av rekryteringsföretagen tycks även ha en policy att inte svara på förfrågningar om information om hur de tjänster som de erbjuder utförs. Endast 10 svarade av de 50 rekryteringsföretag som kontaktades. Dessa utmaningar bidrog till att endast 10 intervjuer kunde genomföras under den tid som undersökningen pågick. Av dessa 10 rekryterar 9 fler än 50 personer per år, och en representerar ett rekryteringsföretag som rekryterar mellan 11–20 personer per år.

Respondenterna var överlag CTO (chief technology officer, det vill säga it-chef eller teknisk direktör) eller motsvarande typ av hr-chef för att säkerställa tillräcklig kunskap för att svara på frågorna. Syftet med dessa intervjuer var att fånga upp hur rekryteringstjänster använder AI och annat automatiserat beslutsfattande och förstår deras risker för diskriminering.

Urvalet tog avstamp från en lista från arbetsgivarorganisationen Kompetensföretagen (en del av Almega) över de 50 största så kallade bemanningsföretagen (som inkluderar rekryteringsföretag) i landet baserat på omsättning. Denna släpps varje kvartal och den senast publicerade valdes. När indikationer på att denna grupp respondenter var särskilt svåra att nå framträdde expanderade uppdragstagarna urvalet till att inkludera fler av de som inte tillhör de 50 största, men det visade sig inte heller leda till att 15 stycken kunde nås.

Genom intervjuer med 10 respondenter från de största rekryteringsbolagen drar vi framför allt kvalitativa slutsatser gällande medvetenhet och kunskapsnivåer kring risker för diskriminering i användning av AI och annat automatiserat beslutsfattande inom rekrytering. Det finns indikationer att peka på, och möjliga förhållningssätt och organisatoriska upplägg går att teckna.

Målgrupp 2: större företag

Gruppen med större svenska arbetsgivare delades upp i två grupper och består dels av 50 intervjuer med företagsrepresentanter för företag med mellan 100–249 anställda, och dels med 50 representanter från arbetsgivare i privat sektor med fler än 250 anställda, varav 23 stycken med mer än 500 anställda, se tabell 1. Denna enkät riktades mot hr-chef eller annan högre chef med tydligt rekryteringsansvar eller motsvarande befattningar.

Tabell 1. Fördelning av urval av större arbetsgivare

Totalt antal intervjuer 100.

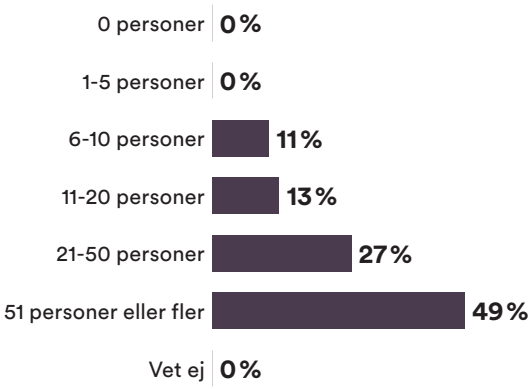
Antal anställda	Andel
100–149 anställda	22 %
150–249 anställda	28 %
250–499 anställda	27 %
500+ anställda	23 %

Storleken på urvalen är anpassade utifrån både projektets budget och insiktsbehov.

Urvalet av större företag representerar en grupp som rekryterar relativt mycket. Omkring hälften (49 procent) av de tillfrågade representerar arbetsgivare som rekryterar mer än 50 personer per år, se figur 1.

Figur 1: Screeningfråga 1. Hur många personer anställer ni ungefär per år?

Antal intervjuer: 100.



Enligt SCB:s statistik finns det omkring 1 904 stycken företag i spannet 100–249 anställda och omkring 1 205 stycken i spannet 250 anställda och uppåt.[1] En bedömning gjordes att 50 intervjuer från varje grupp är tillräckligt stort urval. Detta ska ses som en undersökning av användningen, och dessa antal ger oss en god uppfattning. Av dessa företag rekryterar ungefär hälften (49 procent) fler än 50 personer per år.

Syftet med dessa intervjuer var att få en överblick över användningen av AI och annat automatiserat beslutsfattande vid rekrytering och arbetsledning (management) och fördelning av arbete, och behandlade bland annat:

- anställningsmetoder (till exempel vilka underleverantörer och andra tjänster som används, om det finns någon organisationsintern automatisering eller automatiserad databehandling)
- om datainsamling och automatiserade beslutsstöd används för arbetsledning (till exempel vilka verktyg och metoder som används, omfattning av datainsamling och automatisering, hur viktiga de anses vara för beslutsfattandet)
- vilken bakgrundsinformation de har om kandidater och om de aktivt arbetar med att minska risken för diskriminering i användning av AI och annat automatiserat beslutsfattande.

1 SCB (den 18 januari 2023).

Bilaga 3: Enkät rekrytering/ bemanningsföretag

Förklaringstext:

På uppdrag av Diskrimineringsombudsmannen undersöks här användningen av artificiell intelligens (AI) och annat automatiserat beslutsfattande vid rekryteringsprocesser i arbetslivet.

Automatiserat beslutsfattande omfattar algoritmiska och automatiserade beslutsprocesser såväl med som utan artificiell intelligens. I begreppet inkluderar vi både helt automatiserat beslutsfattande och automatiserade processer som används som beslutstöd.

Screeningfrågor

Hur många personer rekryterar ni ungefär per år?

- 0 → screenas ut
- 1–5 → screenas ut
- 6–10
- 11–20
- 21–50
- 51+
- Vet ej → screenas ut

Enkät

1. I vilken utsträckning inkluderar de system ni använder för att hantera rekrytering olika former av automatiserat beslutsfattande eller automatiserat beslutsstöd?

   - I mycket hög utsträckning
   - I ganska hög utsträckning
   - Varken eller
   - I ganska låg utsträckning
   - I mycket låg utsträckning/inte alls
   - Vet ej

2. I vilken utsträckning inkluderar de tekniska system ni använder för att hantera rekrytering olika former av AI-funktionalitet (exempelvis maskininlärning)?

   - I mycket hög utsträckning
   - I ganska hög utsträckning
   - Varken eller
   - I ganska låg utsträckning
   - I mycket låg utsträckning/inte alls
   - Vet ej

[OM svarat alt d, e eller f i Q2]

3. Tror du att ni kommer att skaffa ett AI-baserat system och annat system för automatiserat beslutstöd för rekrytering inom de kommande tre åren?

   - Ja
   - Nej
   - Vet ej

4. Hur får ni in olika uppgifter om lämpliga kandidater?
Flera svar möjliga
Genom: [roteras]

   a. Vi testar kandidater genom spel, IQ-tester eller personlighetstester
   b. Vi ber dem spela in en video med svar, som vi analyserar
   c. Från egna webb-formulär
   d. Webbanvändningsdata inklusive cookies från den sökande på er egen hemsida
   e. Från sociala medieprofiler (till exempel Facebook, LinkedIn, Instagram, TikTok, Twitter med flera)
   f. Olika befintliga plattformar, exempelvis på Arbetsförmedlingen, Monster et cetera
   g. Från datamäklare som samlat på sig profiler genom olika sorters datainsamling, och erbjuder till försäljning
   h. Annat, nämligen
   i. Vet ej

5. Använder ni AI-system eller andra verktyg för automatiserat beslutsfattande för att ...
Flera svar är möjliga

   a. Matcha aktiva arbetssökande med jobb utifrån deras professionella profil?
   b. Matcha kandidater (både aktiva och passiva) med riktad annonsering mot deras professionella profil?
   c. Sortera eller ranka sökande för arbetsgivare?
   d. Annat, nämligen:
   e. Använder ej
   f. Vet ej

6. Vilka av följande bakgrundsdata har ni tillgång till i de databaser ni använder er av när ni rekryterar?
   Flera svar är möjliga

   a. Ålder
   b. Kön
   c. Utbildning
   d. Språkkunskaper
   e. Födelseland
   f. Medborgarskap
   g. Postnummer
   h. Eventuell funktionsnedsättning
   i. Sexuell läggning
   j. Etnisk tillhörighet
   k. Religion eller annan trosuppfattning
   l. Utseende
   m. Inga bakgrundsdata
   n. Annat, nämligen:
   o. Vet ej

[Svarat alt d-l i Q6]

7. Hur används dessa uppgifter?

   - Öppen:

[om i mycket hög utsträckning, i ganska hög utsträckning i Q2]

8. Känner du till hur ert AI-system har tränats för att hitta rätt kandidater, exempelvis gällande själva träningsdatan för AI-systemet?

   a. Ja

   b. Nej

   c. Vet ej

[om i mycket hög utsträckning, i ganska hög utsträckning i Q2]

9. Känner du till om systemleverantören har medvetenhet om risker för diskriminering?

   a. Ja

   b. Nej

   c. Vet ej

När vi frågar om diskriminering så menar vi diskriminering utifrån diskrimineringslagens diskrimineringsgrunder. Dessa är kön, könsöverskridande identitet eller uttryck, etnisk tillhörighet, religion eller annan trosuppfattning, funktionsnedsättning, sexuell läggning och ålder.

Om alt a – d i Q1
Om alt a – d i Q2
Om alt a – e i Q5

10. Förebygger ni diskriminering i er användning av AI eller andra former av automatiserat beslutsfattande? Flera svar möjliga

    a. Ja, vi har interna riktlinjer och rutiner för att förebygga diskriminerande effekter

    b. Ja, vi kvalitetssäkrar våra system för att förebygga diskriminering

    c. Ja, vi har utvärderat våra system för att säkerställa överensstämmelse med diskrimineringslagen

    d. Ja, vi har låtit en extern tredje part (specialiserad på diskrimineringsfrågor) granska vår användning av dessa tjänster/produkter

    e. Ja, vi har metoder för att motverka så kallad automation bias, det vill säga för att förhindra att beslut fattas alltför okritiskt i förhållande till beslutsstöd

    f. Nej, vi har inte tänkt på detta

    g. Nej, men vi litar på att leverantören har säkrat systemet mot diskriminering

    h. Använder ej

    i. Annat, nämligen:

    j. Vet ej

[Om d i Q10]

11. Ge gärna något exempel på en extern tredje part (till exempel en konsult specialiserad mot diskrimineringsfrågor) ni har använt er av!

    - Öppen:

12. Vilket stöd/kunskapsbehov har ni för att förebygga risker för diskriminering i AI-system eller system för automatiserat beslutsfattande eller automatiserat beslutsstöd?

    - Öppen:

Om svarat alt: a, b, c eller d i Q1

13. Tycker du att man bör informera en kandidat om att man använder ett AI-system och annat automatiserat beslutsfattande för att sålla fram rätt kandidat?

- Ja, och vi gör det i dag
- Ja, men vi gör inte det i dag
- Nej
- Vet ej

Får vi kontakta dig igen, om ytterligare frågor dyker upp?

- Ja (NAMN + TELEFONNUMMER)
- Nej

Bilaga 4: Enkät till större svenska arbetsgivare

Screeningfrågor

1. Hur många personer anställer ni ungefär per år?
    a. 0 → screenas ut
    b. 1–5 → screenas ut
    c. 6–10
    d. 11–20
    e. 21–50
    f. 51+
    g. Vet ej → screenas ut

2. I vilken utsträckning inkluderar de system ni använder för att hantera rekrytering olika former av automatiserat beslutsfattande eller automatiserat beslutsstöd?
    a. I mycket hög utsträckning
    b. I ganska hög utsträckning
    c. Varken eller
    d. I ganska låg utsträckning
    e. I mycket låg utsträckning/inte alls
    f. Vet ej

3. I vilken utsträckning använder ni er av olika former av automatiserat beslutsfattande eller automatiserat beslutsstöd i arbetsledning som har med medarbetare att göra?

   a. I mycket hög utsträckning
   b. I ganska hög utsträckning
   c. Varken eller
   d. I ganska låg utsträckning
   e. I mycket låg utsträckning/inte alls
   f. Vet ej

4. Det finns olika tekniska lösningar som ibland används i rekryteringsprocesser. I era rekryteringsprocesser, använder ni er av något av följande: Flera svar är möjliga

   a. En rekryteringsplattform (till exempel Reachmee, Varbi, LinkedIn Recruiter, Teamtailor)
   b. Automatiserad filtrering, sortering eller poängsättning av cv:s
   c. Sociala medier (till exempel Facebook, LinkedIn) för att matcha kandidater med jobbet
   d. Sökmotorer (till exempel Google) eller sociala medier (till exempel Facebook, LinkedIn) för att samla in information om kandidater (kallas ibland för cybervetting)
   e. Ni outsourcar processerna till ett rekryteringsföretag
   f. Annat, nämligen:
   g. Använder ej

5. Det finns olika tekniska lösningar som ibland används för personalledning. I er personalledning, gör ni något av följande: Flera svar är möjliga

   a. Samlar in kvantitativa data om anställdas prestationer, exempelvis med hjälp av program som Deel, Papaya, Microsoft Teams ”employee monitoring”

   b. Använder automatisk poängsättning för att underlätta utvärdering av prestationer

   c. Använder automatiserad filtrering, sortering eller poängsättning för att underlätta lönesättning och förändring av andra anställningsvillkor, befordran, utbildning och övrig kompetensutveckling

   d. Använder automatiserad filtrering, sortering eller poängsättning för omplacering eller grund för uppsägning

   e. Automatiserar ni skiftschemaläggning eller uppgiftstilldelning

   f. Annat, nämligen:

   g. Använder ej

6. Vid en normal rekrytering: vilken av följande bakgrundsdata om era sökande brukar ni ha tillgång till?

   a. Ålder
   b. Kön
   c. Utbildning
   d. Språkkunskaper
   e. Födelseland
   f. Medborgarskap
   g. Postnummer
   h. Eventuell funktionsnedsättning
   i. Sexuell läggning
   j. Etnisk tillhörighet
   k. Religion eller annan trosuppfattning
   l. Utseende
   m. Inga bakgrundsdata
   n. Annat, nämligen
   o. Vet ej

[Om alt d-k på Q5]

7. Hur används dessa uppgifter?

   - Öppen

[Om alt: a, b, c eller d i Q1]
Om alt: a – f i Q4
Om alt: a – d + f i Q3

När vi frågar om diskriminering så menar vi diskriminering utifrån diskrimineringslagens diskrimineringsgrunder. Dessa är kön, könsöverskridande identitet eller uttryck, etnisk tillhörighet, religion eller annan trosuppfattning, funktionsnedsättning, sexuell läggning och ålder.

8. Förebygger ni aktivt diskriminering i er användning av AI eller andra former av automatiserat beslutsfattande som arbetsgivare? Flera svar är möjliga

- **a.** Ja, vi har interna riktlinjer och rutiner för att förebygga diskriminerande effekter
- **b.** Ja, vi kvalitetssäkrar våra system för att förebygga diskriminerande effekter
- **c.** Ja, vi har utvärderat våra system för att säkerställa överensstämmelse med diskrimineringslagen
- **d.** Ja, vi har låtit en extern tredje part (specialiserad på diskrimineringsfrågor) granska vår användning av dessa tjänster/produkter
- **e.** Ja, vi har metoder för att motverka så kallad automation bias, det vill säga för att förhindra att beslut fattas alltför okritiskt i förhållande till beslutsstöd
- **f.** Nej, vi har inte tänkt på detta
- **g.** Nej, men vi litar på att leverantören har säkrat systemet mot diskriminering
- **h.** Annat, nämligen:
- **i.** Vet ej

[Om alt d i Q7]

9. Ge gärna något exempel på en extern tredje part (till exempel en konsult specialiserad mot diskrimineringsfrågor) ni har använt er av!
   - Öppen:

10. Om ert företag har upphandlat ett AI-system eller liknande, har ni säkerställt att det inte är diskriminerande?
    a. Ja
    b. Nej
    c. Vet ej
    d. Har ej upphandlat ett AI-system eller liknande

[Om alt a (JA) i Q9]

11. Hur har ni säkerställt att det inte är diskriminerande?
    - Öppen:

12. Vem tror du bär huvudansvaret för diskriminering som uppstår som en konsekvens av användningen av ett AI-system och annat automatiserat beslutsfattande som används av ett företag?
    a. Systemleverantören/systemutvecklaren
    b. Företaget som använder systemet
    c. Annan, nämligen:
    d. Vet ej

Om svarat alt: a, b, c eller d i Q1

13. Tycker du att man bör informera en kandidat om att man använder ett AI-system och annat automatiserat beslutsfattande för att sålla fram rätt kandidat?

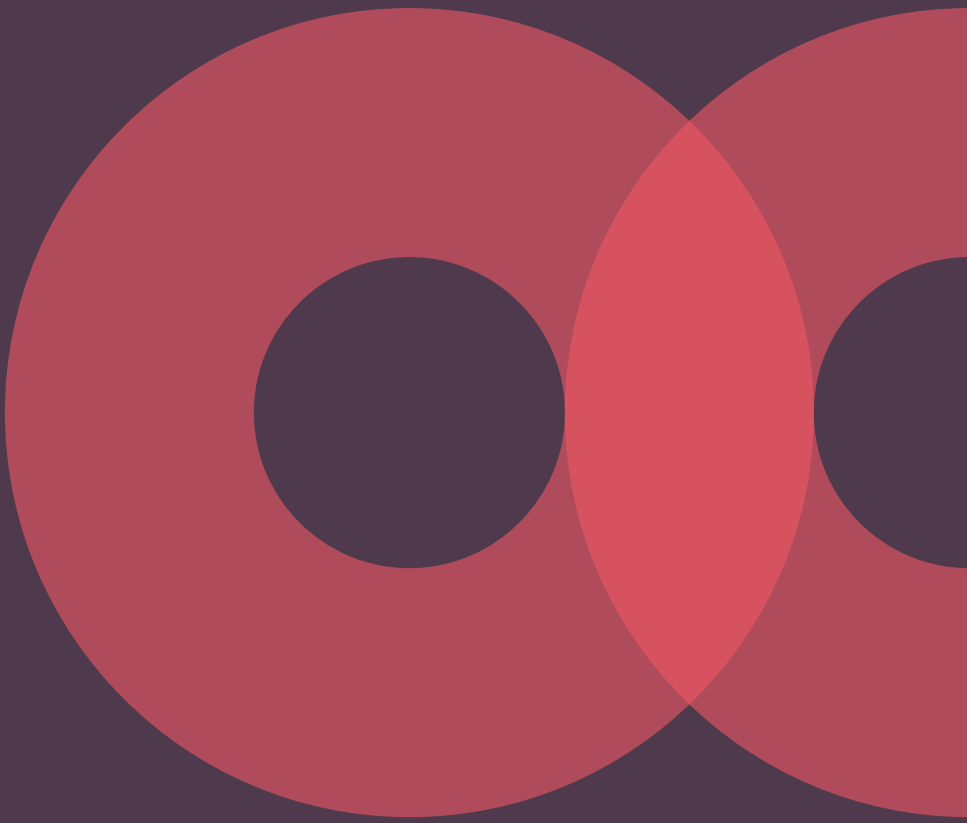
- **a.** Ja, och vi gör det i dag
- **b.** Ja, men vi gör inte det i dag
- **c.** Nej
- **d.** Vet ej

14. Vilket stöd/kunskapsbehov har ni för att förebygga risker för diskriminering i AI-system och annat automatiserat beslutsfattande?

- Öppen:

Får vi kontakta dig igen, om ytterligare frågor dyker upp?

- Ja (NAMN + TELEFONNUMMER)
- Nej



DO
www.do.se
Box 4057
169 04 Solna
Telefon 08-120 20 700
Facebook/Diskrimineringsombudsmannen
Linkedin/Diskrimineringsombudsmannen, DO
X/@DO_Sverige

ISBN 978-91-88175-48-9