Gemeente Amsterdam

Versie 0.1
Februari 2023

Fairness analyse 'Blurring as a Service'

Computer Vision Team
Directie Digitalisering & Innovatie
Gemeente Amsterdam

Voorwoord

Gemeenten werken steeds beeldgerichter. Scanauto's brengen de openbare ruimte in kaart om de dienstverlening van de gemeente te verbeteren, burgers delen beelden met de gemeente bij vergunningsaanvragen of bij meldingen in de openbare ruimte. Gemeenteambtenaren leggen voor en na situaties vast op beeld ter ondersteuning van de werkzaamheden.

De gemeente heeft daarom meerdere beelden in bezit en deze beelden kunnen persoonsgegevens bevatten, zoals personen en kentekens. Daarmee moeten gemeenten maatregelen nemen die voortkomen uit de Europese privacy wet, de Algemene Verordening Gegevensbescherming (AVG), waaronder het anonimiseren van beelden.

Het Computer Vision Team van de Gemeente Amsterdam heeft in samenwerking met Gemeente Utrecht in het kader van het innovatiebudget van BZK een bluralgoritme ontwikkeld. Dit algoritme is onderdeel van 'Blurring as a Service' dat aangeboden kan worden aan verschillende overheidsinstanties zodat de overheid minder afhankelijk wordt van marktpartijen voor een cruciale privacymaatregel, namelijk het anonimiseren van beelden van de openbare ruimte.

Marktpartijen bepalen op dit moment de norm als het gaat om de manier waarop en de kwaliteit waarmee beelden geanonimiseerd worden. Gemeenten stellen hoge eisen aan het anonimiseren van beelden en zien deze eisen op dit moment niet vervuld in de markt.
'Blurring as a Service' levert een transparante en kwalitatief hoogwaardige anonimisering van beelden die bovenal *eerlijk* is. Dat wil zeggen ; *elke inwoner of bezoeker van een Nederlandse gemeente moet een gelijke kans krijgen om geanonimiseerd te worden*.

Dit document beschrijft de aanpak en analyse om deze zgn. fairness te kunnen garanderen.

Inhoud

1.Definitie fairness 4

1.1Type risico's in AAI-algoritmen kennen verschillende risico's en een samenvatting daarvan is hieronder weergegeven 4

1.2 Fairness categorieën 5

1.1 Fairness metriek 6

2.Collectie Data 8

2.1 Selectie beelden trainingset 8

2.2 Annotatie 8

3. Resultaten Fairness 9

11

1.Definitie fairness

1.1Type risico's in A

AI-algoritmen kennen verschillende risico's en een samenvatting daarvan is hieronder

weergegeven

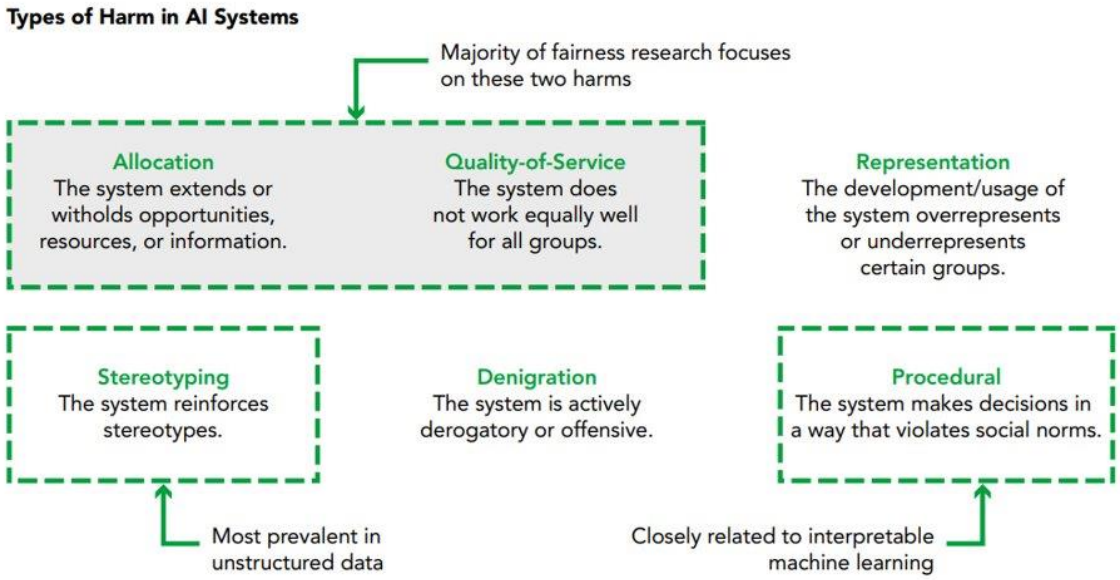


Figure 5: A summary of the harms that can prevail in AI systems. Source: Hilde Weerts

Bij 'Blurring as a Service' gaat het om een beeldherkenningsalgoritmen waarmee personen en kentekens herkend moeten worden, zodat deze geanonimiseerd kunnen worden. Van bovenstaande risico's worden, in volgorde van belangrijkheid en waarschijnlijkheid, de volgende risico's relevant gezien voor deze service:

1. **Quality-of-Service**

2. **Representation**
3. **Denigration**

Er worden geen beslissingen genomen op basis van het algoritme en daarmee worden de risico's Allocation, Stereotyping en Procedural als niet relevant beschouwd.

1.2 Fairness categorieën

Uitgangspunt voor het in kaart brengen van alle fairness categorieën is de Nederlandse en Europese wetgeving. Hieronder een overzicht van alle beschermde attributen op basis waarvan geen discriminatie is toegestaan.



Figure 1: The protected attributes on which it is highly undesired or prohibited to discriminate against

Aangezien 'Blurring as a Service' en beeldherkenningsalgoritme betreft, kan discriminatie enkel plaatsvinden op basis van visuele kenmerken. Vandaar dat bovenstaande lijst is vertaald naar visuele kenmerken.

Type	Visueel kenmerk	Fairness categorie	Subcategorieën
Persoon	Geslacht/Sex	geslacht/sex	Man / Vrouw / Onbekend
Persoon	Leeftijd	leeftijd	Kind / Volwassen / Bejaard / Onbekend
Persoon	Huidskleur	huidskleur, nationaliteit, migratie achtergrond	Lichter, midden, donkerder, onbekend

Perso on	Kleding	etniciteit, sociale klasse, beroep, religie	NVT
Perso on	Attributen	invaliditeit, beroep, interesse/hobby, nationaliteit	NVT

De groepen zijn onderverdeeld in de context van beeldherkenning. In andere contexten zouden het zeker denkbaar zijn dat andere groepen gemaakt zouden worden. Het gaat hier puur om oppervlakkig zichtbare verschillen, zoals dat een kind kleiner is dan een volwassenene.

Voor elke categorie, waar het mogelijk was om het in groepen onder te verdelen, zijn subcategorieën gebruikt. Hiervoor is het Fairness handboek pagina 6 gebruikt.

Op basis van deze visuele kenmerken mag het algoritme niet discrimineren. Dat wil zeggen; elke persoon of kenteken die onderdeel uit maakt van deze verschillende categorieën moet een gelijke kans hebben om geanonimiseerd te worden.

1.1 Fairness metriek

Om een metric voor fairness te bepalen, is de onderstaande beslisboom gebruikt. Deze beslisboom komt uit het Fairness handboek dat ontwikkeld is door de Advanced Analytics Team van de Gemeente Amsterdam.

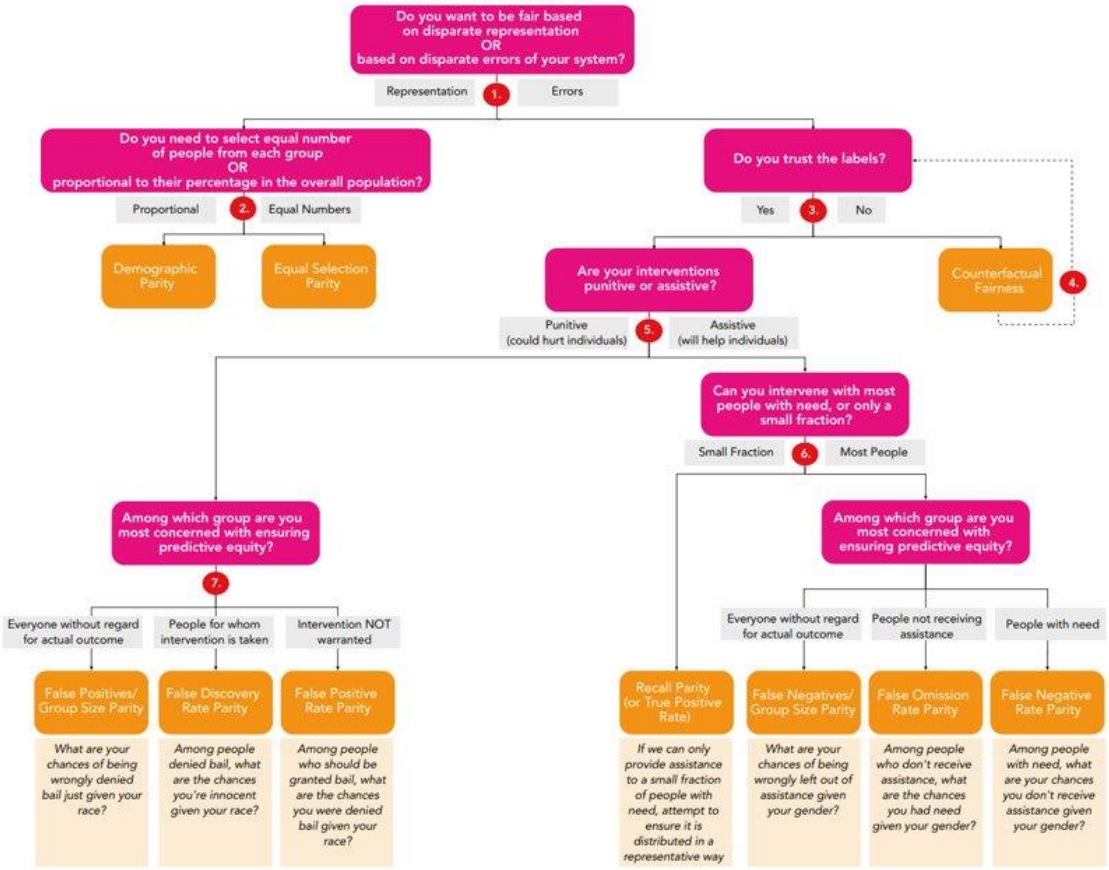
Antwoord op vraag 1:
Voor het blur algoritme zijn we geïnteresseerd of het algoritme gelijk is wat betreft de fouten. Het zou niet zo moeten zijnen dat een bepaalde groep niet wordt geblurd.

Antwoord op vraag 3:
Ja, we vertrouwen de labels aangezien, die intern met eigen ontwikkelde instructie geplaatst zijn.

Antwoord op vraag 5:
Het systeem is bedoeld om de burger te helpen door hun anonimiteit te beschermen.

Antwoord op vraag 6:
Iedereen in de stad zou geblurd moeten zijn en de meeste daarvan kunnen we helpen

De metric, die is gebruikt, is de F*alse Negative Rate* omdat het, zoals in de beschrijving in de beslisboom, helpt bij het geven van inzicht in de kans dat een persoon of kentekenplaat niet herkend wordt gegeven de groep waartoe je behoort.


Do you want to be fair based on disparate representation OR based on disparate errors of your system?
Representation
1.
Errors
Do you need to select equal number of people from each group OR proportional to their percentage in the overall population?
Proportional
2.
Equal Numbers
Demographic Parity
Equal Selection Parity
Do you trust the labels?
Yes
3.
No
Are your interventions punitive or assistive?
Counterfactual Fairness
4.
Punitive (could hurt individuals)
5.
Assistive (will help individuals)
Can you intervene with most people with need, or only a small fraction?
Small Fraction
6.
Most People
Among which group are you most concerned with ensuring predictive equity?
7.
Everyone without regard for actual outcome
People for whom intervention is taken
Intervention NOT warranted
False Positives/ Group Size Parity
False Discovery Rate Parity
False Positive Rate Parity
What are your chances of being wrongly denied bail just given your race?
Among people denied bail, what are the chances you're innocent given your race?
Among people who should be granted bail, what are the chances you were denied bail given your race?
Among which group are you most concerned with ensuring predictive equity?
Everyone without regard for actual outcome
People not receiving assistance
People with need
Recall Parity (or True Positive Rate)
False Negatives/ Group Size Parity
False Omission Rate Parity
False Negative Rate Parity
If we can only provide assistance to a small fraction of people with need, attempt to ensure it is distributed in a representative way
What are your chances of being wrongly left out of assistance given your gender?
Among people who don't receive assistance, what are the chances you had need given your gender?
Among people with need, what are your chances you don't receive assistance given your gender?

2.Collectie Data

2.1 Selectie beelden trainingset

Het doel bij het verzamelen van de data is om een diverse dataset te verkrijgen. Het getrainde model moet bekend worden met alle stadsbeelden, denk bijvoorbeeld aan parken, grachten, flats, maar ook met alle soorten mensen, zoals van verschillende leeftijden en huidskleur.

Ook is het belangrijk dat er in de dataset een diversiteit aan de mensen en kentekenplaten aanwezig is om het model met de verscheidenheid bekend mee te maken. In onze definitie van *fair*ness (zie fairness document) zijn de verschillende groepen beschreven. Aan de hand van deze groepen zijn locatie bepaald, bijvoorbeeld om kinderen in de dataset te hebben, zijn er beelden *gesampled* rondom scholen.

In het fairness Excel bestand, is de lijst met locaties beschreven waar de beelden zijn verzameld en om welke reden die locatie is gekozen

2.2 Annotatie

Voor het annoteren van de beelden is er onderscheid gemaakt tussen de trainingsdata en de andere sets. In de trainingsdata zijn alle beelden gelabeld met het persoon of kentekenplaat klasse. Voor de test en validatie is er ook geannoteerd op de belangrijk fairness categorieën, zodat we hier metingen voor kunnen doen in de evaluatie. Zo kan er bijvoorbeeld gemeten worden hoeveel personen het model mist met een donkere huidskleur ten opzichte van mensen met een lichtere huidskleur.

Het annotatie protocol staat ook beschreven in het fairness Excel bestand.

3. Resultaten Fairness

In dit hoofdstuk worden de resultaten voor de fairness getoond. Voor alle beschreven groepen is een evaluatie gedaan om het meetbaar te kunnen maken over een bias bestaat voor één van de groepen.

Het is goed om te vermelden dat er weinig wetenschappelijk literatuur is te vinden over vergelijkbaar werk en het lijkt alsof wij hier een eerste in zijn.

In tabel 6 staan de resultaten voor de verschillende huidskleuren. Hoe lager de False Negative Rate des te beter. Het eerste wat opvalt is dat de scores heel dicht bij elkaar liggen en dat er nauwelijks onderscheid is performance van het model voor verschillende huidskleuren. Er blijven natuurlijk altijd verschillen en sommige personen zullen in het algemeen lastiger te herkennen zijn dan anderen.

PERSON / SKIN TONE

	False Negative Rate
Lighter	0.533
Medium	0.45
Darker	0.487
Unknown	0.471

Tabel 6: De False Negatives Rates voor de groep personen met verschillende huidstonen op de validatieset.

In Tabel 7 worden de resultaten getoond voor verschillende groottes. Hier valt te zien dat de scores voor personen die dicht bij de camera veel hoger zijn, dan persoon die ver weg staan. Ook hier zien we nauwelijks verschil in resultaat tussen de verschillende groepen en kunnen we concluderen dat het model fair oordeelt.

PERSON / SKIN TONE

	False Negative Rate Small	**False Negative Rate Medium**	**False Negative Rate Large**
Lighter	0.934	0.5	0.277
Medium	0.8	0.396	0.23
Darker	0.856	0.565	0.149
Unknown	0.818	0.346	0.154

Tabel 7: De False Negatives Rates voor de groep personen met verschillende huidstonen per grootte op de validatieset.

In Tabel 8 worden de resultaten voor de verschillende leeftijden gezien. Hier wordt een wat groter verschil in resultaten geconstateerd voor kinderen. Het is belangrijk om te vermelden dat er maar 10 kinderen in de validatie zaten en dat de groep wellicht niet goed genoeg gerepresenteerd was. Daarbij blijkt uit tabel 9 dat kinderen die dicht bij de camera wel goed herkend worden en dat het vooral over kinderen gaat die ver van de camera staan. Om dit beter te kunnen evalueren, wordt er in toekomst meer kinderen aan de validatie toegevoegd.

PERSON / AGE

	False Negative Rate
Children	0.688
Adult	0.486
Eldery	0.442
Unknown	1.0

Tabel 8: De False Negatives Rates voor de groep personen met verschillende leeftijden op de validatieset.

Person / AGE

	False Negative Rate Small	**False Negative Rate Medium**	**False Negative Rate Large**
Children	0.933	0.778	0.125
Adult	0.864	0.467	0.22
Eldery	0.812	0.375	0.25
Unknown	1.0	NVT	NVT

Tabel 9: De False Negatives Rates voor de groep personen met verschillende leeftijden per grootte op de validatieset.

In tabel 10 en 11 zien we opnieuw vergelijkbare resultaten voor verschillende gender. Er lijkt door het model geen onderscheid te worden gemaakt tussen de verschillende groepen.

PERSON / GENDER

	False Negative Rate
Male	0.503
Female	0.457
Unknown	0.692

Tabel 10: De False Negatives Rates voor de groep personen met verschillende gender op de validatieset.

	False Negative Rate Small	**False Negative Rate Medium**	**False Negative Rate Large**
Male	0.881	0.505	0.24
Female	0.836	0.439	0.193
Unknown	0.917	0.5	0.222

Tabel 11: De False Negatives Rates voor de groep personen met verschillende gender per grootte op de validatieset.

Concluderend kunnen we stellen dat de bias geminimaliseerd is en dat het model nauwelijks tot niet discrimineert op het gebied van gender, leeftijd en huidskleur. Uit de gebreide analyse en metingen zien we enkel een aanzienlijk verschil voor de groep kinderen, al lijkt het er sterk op dat er niet genoeg samples van deze groep in de evaluatie zitten. In de toekomst zullen we dit uitgebreider meten.