Automated Decision-Making Fairness in an AI-driven World

Araujo, T.B.; de Vreese, C.H.; Helberger, N.; Kruikemeier, S.; van Weert, J.C.M.; Bol, N.; Oberski, Daniel; Pechenizkiy, Mykola; Schaap, Gabi; Taylor, Linnet

Digital Communication Methods Lab
RPA Communication
University of Amsterdam

Automated Decision-Making Fairness in an AI-driven World:

Public Perceptions, Hopes and Concerns

Key findings

- Respondents report generally low knowledge levels concerning artificial intelligence (AI) and algorithms;
- AI leading to manipulation, risk or unacceptable outcomes among the highest scoring public *concerns*, whereas usefulness and potential for objective treatment are highlighted as *opportunities*;
- Health, Justice, Commerce, and Media & Politics are the most frequently mentioned examples that respondents provide when considering automated decision-making (ADM) by AI;
- Human Control, Human Dignity, Fairness and Accuracy are key values prioritised by respondents when considering ADM by AI.

Theo Araujo[1], Claes de Vreese[1], Natali Helberger[1], Sanne Kruikemeier[1], Julia van Weert[1], Nadine Bol[1], Daniel Oberski[2], Mykola Pechenizkiy[3], Gabi Schaap[4] and Linnet Taylor[5]

[1]University of Amsterdam
[2]Utrecht University
[3]Eindhoven University of Technology
[4]Radboud University
[5]Tilburg University

25 September, 2018

Table of contents

Introduction 3
Knowledge 4
Perceived importance 6
Hopes and Concerns 8
Sectors associated with ADM by AI 10
Values respondents find important for ADM by AI 12
About this report 15
References 20

Introduction

Ongoing advances in artificial intelligence (AI) are increasingly part of scientific efforts as well as the public debate and the media agenda, raising hopes and concerns about the impact of automated decision-making across different sectors of our society. This topic is receiving increasing attention at both national and cross-national levels.

In the Netherlands, for example, the Dutch National Research Agenda [1] highlights the need to investigate the impact of these new technologies on humans and society. Automated decision-making is also highlighted by the Association of Dutch Universities' (VSNU) Digital Society Research Agenda [2], including the need to research autonomous systems and automated decision-making (ADM), and by the Digital Agenda from the Dutch government [3], which calls for the investigation of the effects of digitalisation and usage of intelligent technologies across several sectors, including health.

Accelerating research in artificial intelligence, including the development of ethical guidelines, was set as a priority by the European Commission in 2018 [4]. Moreover, challenges brought by big data and algorithms have been highlighted by several reports, including by the White House (e.g., [5]). The same concern is shared by scientists across the world, including recent recommendations issued by the members of ACM Europe [6], one of the top associations in the field.

The present report contributes to informing this public debate, providing the results of a survey with 958 participants recruited from high-quality sample of the Dutch population. The following pages present an overview of public knowledge, perceptions, hopes and concerns about the adoption of AI and ADM across different societal sectors in the Netherlands.

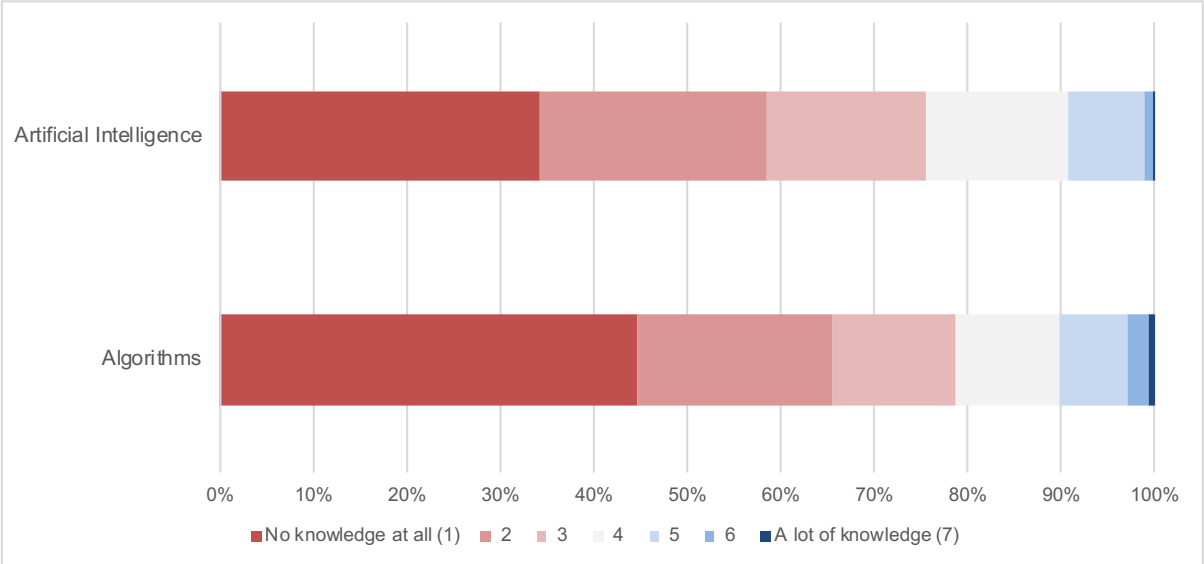
This report is part of a research collaboration between the University of Amsterdam, Tilburg University, Eindhoven University of Technology, Utrecht University and Radboud University on automated decision-making, and forms input to the groups' research on fairness in automated decision-making.

Knowledge

Majority of respondents reports low knowledge about AI and algorithms

Respondents provided very low scores when asked to evaluate their own knowledge about both AI and algorithms, which are key concepts associated with automated decision-making. The average score for knowledge about AI was 2.43 (*SD* = 1.37) on a 7-point scale, with around **75%** of respondents providing scores lower than 4 (middle of the scale). For algorithms, the average was 2.25 (*SD* = 1.47), with almost **79%** of respondents providing low scores for their own knowledge about the topic. Moreover, approximately **34%** of respondents indicated having *no knowledge at all* about artificial intelligence; this rose to **45%** when asked about algorithms.

Figure 1. Self-reported knowledge about AI and algorithms

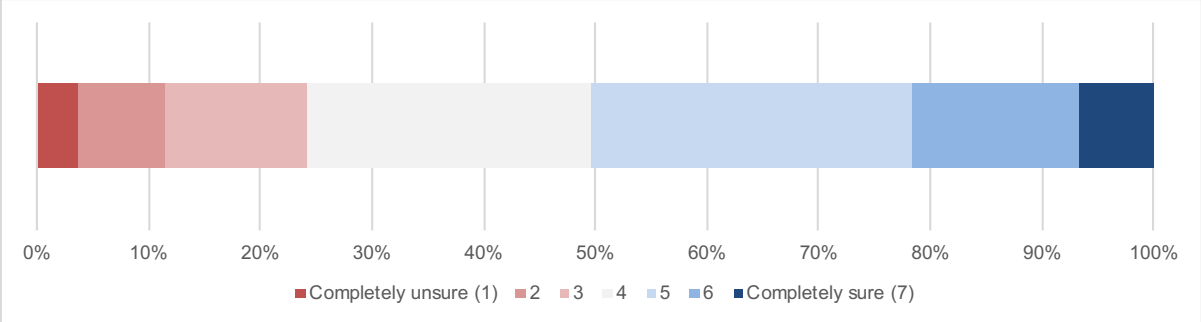


Most respondents were able to give some explanation for ADM

We also asked respondents to explain in their own words what they understood automated decision-making by artificial intelligence or computers to be. Before answering this and the following questions, participants were informed that ADM was defined as “artificial intelligence or computers that take decisions based on data, without involving humans”.

Approximately 25% were unable to provide any explanation. Those who did provide an explanation (N = 717) were moderately confident about how accurate their explanation was, with their average level of confidence being 4.39 (*SD* = 1.45) on a 7-point scale.

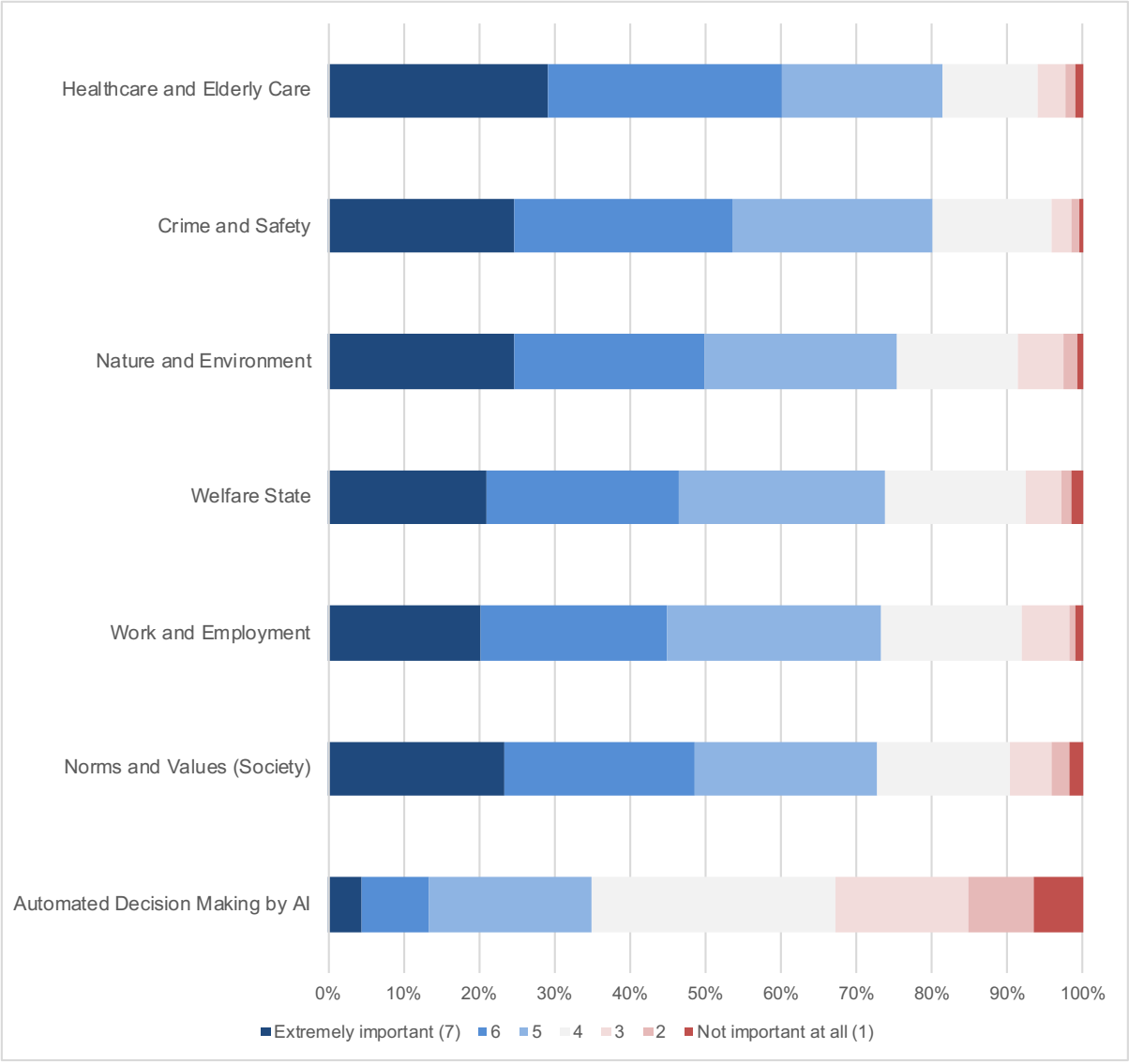
Figure 2. Confidence in own explanation about ADM



Perceived importance

Respondents also indicated the extent to which they believed that ADM by AI would be an important societal trend for the next 5-10 years. For comparison purposes, respondents also provided evaluations concerning a set of randomly selected trends and concerns mentioned in the Citizen Perspectives (Burgerperspectieven) report [7] by the Netherlands Institute for Social Research (Sociaal en Cultureel Planbureau).

Figure 3. Perceived importance of ADM by AI and other societal concerns



The relatively low scores for ADM by AI (*M* = 3.98, *SD* = 1.44)[1] as such, compared to other topics such as Healthcare and Elderly Care (*M* = 5.61, *SD* = 1.28) or Crime and Safety (*M* = 5.52, *SD* = 1.20) seem to indicate that respondents might not consider this topic as a *separate* category or cause of concern. This is likely due to the relative novelty in the public debate and low levels of awareness about ADM by AI compared to the other trends. That being said, trends that will likely be impacted by ADM by AI seem to feature in highly in the set of concerns, and approximately 35% of respondents already indicated that ADM by AI is an important trend (above the mid-point of the scale) for the near future.

[1] All means and standard deviations mentioned in this report are for the complete valid sample (N = 958) unless otherwise noted in the text.

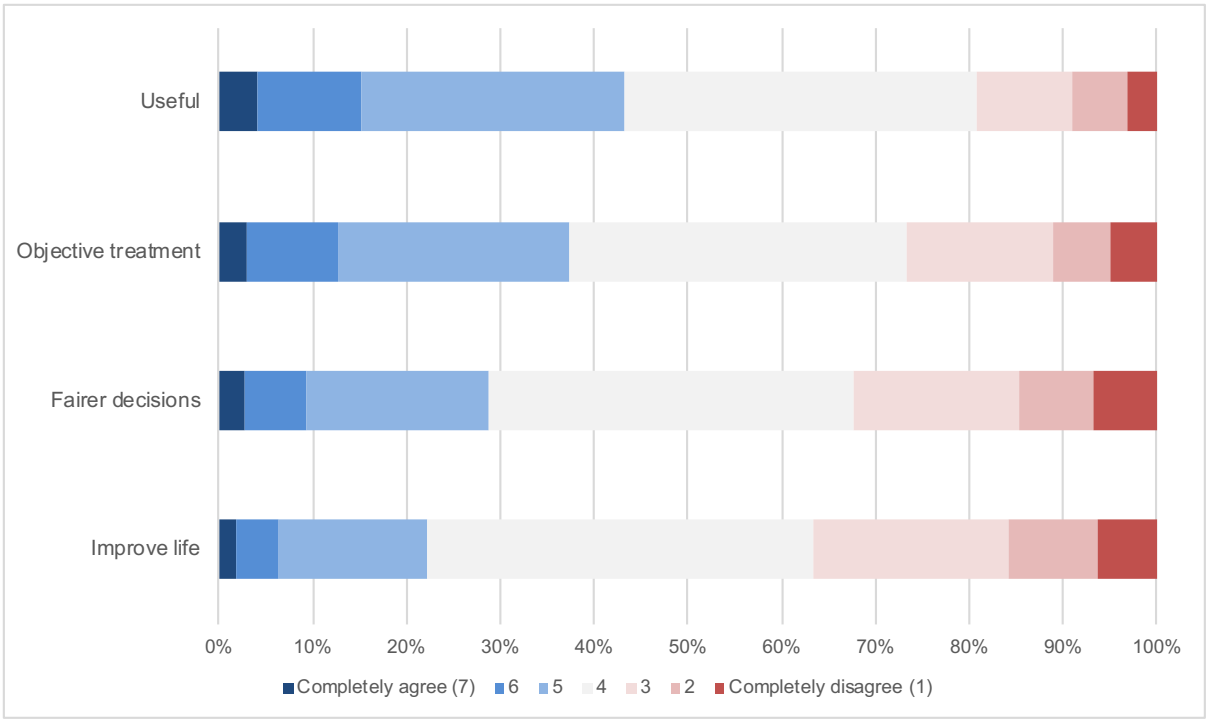
Hopes and Concerns

Respondents were asked to indicate the extent to which they believed ADM and AI would present positive and/ or negative challenges for the society as a whole[2]. The picture that emerged shows that concerns seem to outweigh the positive potential of ADM by AI at present.

ADM seen as potentially useful, but less as fair or improving life

On the one hand, respondents seemed cautiously optimistic about ADM, at least when it comes to its potential usefulness: over 40% of respondents indicated some level of agreement (above the mid-point of the scale) with the statement that ADM by AI would be **useful** (*M* = 4.31, *SD* = 1.27), whereas less than 20% expressed some level of disagreement with the statement. Moreover, about 37% of respondents agreed at some level that ADM by AI would lead to **objective treatment** of people (*M* = 4.10, *SD* = 1.32), with around 27% disagreeing at some level.

Figure 4. Hopes and the positive potential of ADM by AI



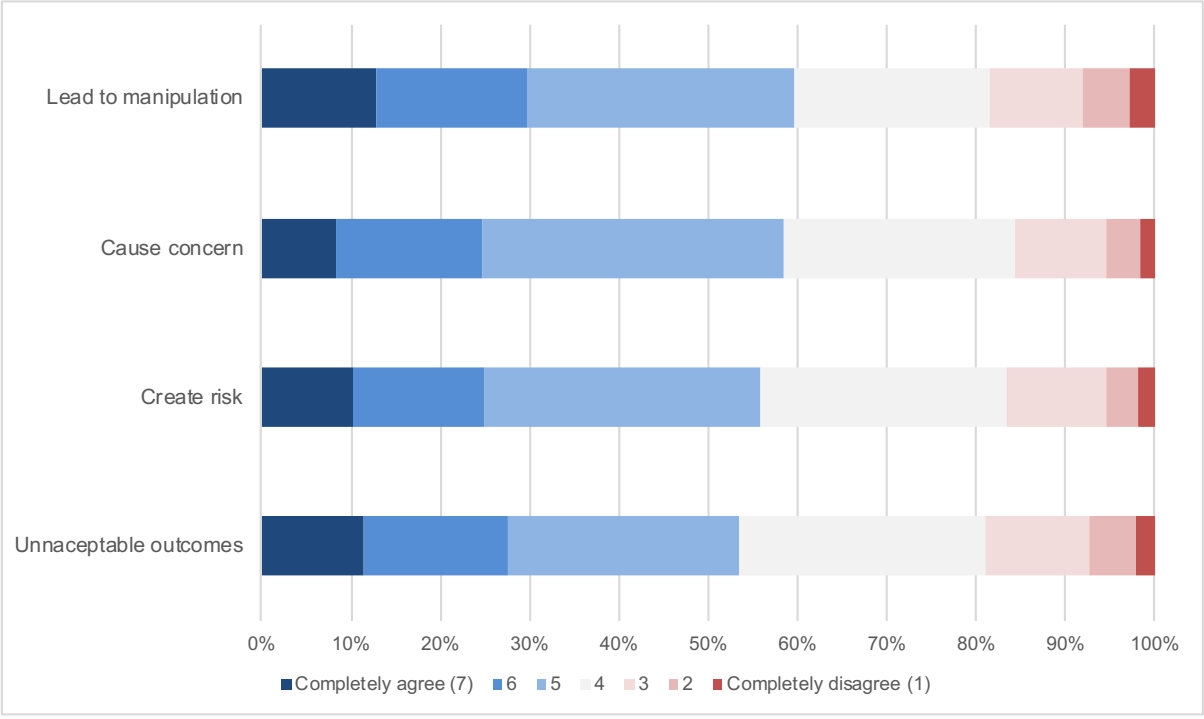
[2] Respondents provided evaluations to a set of 15 statements with positive and negative potential outcomes for ADM by AI based on earlier questionnaires in academic research aimed at measuring benefits, risk, and concerns about AI or new technologies. This section reports the results of the four positive statements and four negative statements with the highest levels of agreement (above mid-point of the scale). Items with very similar wording or that had their wording reversed to test reliability of the scales were excluded from this report.

On the other hand, when it comes to the potential of ADM by AI to improve life or lead to fairer decisions, respondents were much less optimistic. While almost 30% expressed some level of agreement with the notion that that it would lead to **fairer decisions** (*M* = 3.87, *SD* = 1.34), about 32% disagreed at some level with the statement. When it comes to ADM by AI **improving life** (*M* = 3.72, *SD* = 1.26), only around 22% indicated some level of agreement, whereas almost 37% expressed some level of disagreement (below the mid-point of the scale).

Concerns about manipulation and unacceptable results

Along the same lines, respondents were much more in agreement with the potential concerns that ADM by AI may bring. Almost 60% expressed some level of agreement with the statement that ADM may **lead to manipulation** (*M* = 4.73, *SD* = 1.47) or **cause worry** (*M* = 4.69, *SD* = 1.29). Moreover, over 55% expressed some level of support with the notion that ADM by AI may **create risk** (*M* = 4.67, *SD* = 1.34) and around 53% believed that it might outright **lead to unacceptable outcomes** (*M* = 4.64, *SD* = 1.43).

Figure 5. Concerns about the negative potential of ADM by AI



Sectors associated with ADM by AI

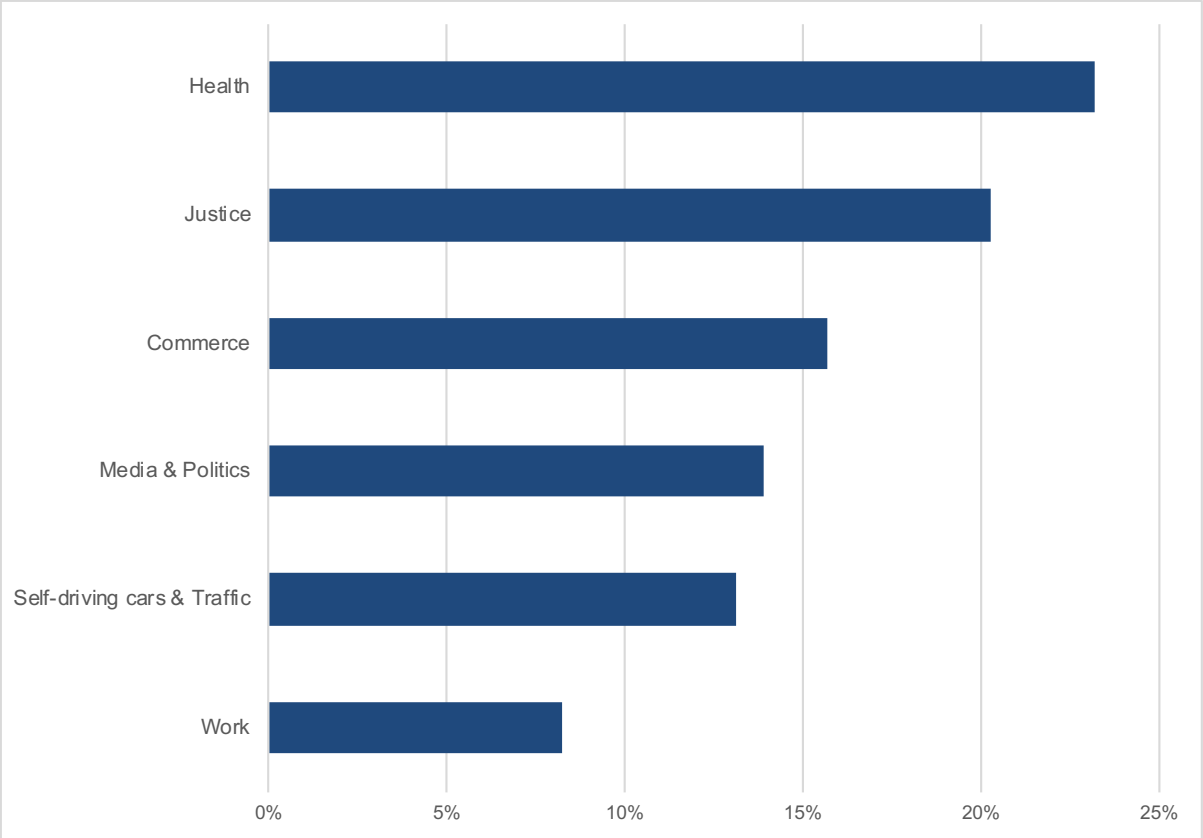
While evaluating the positive and negative potential for ADM by AI, respondents were also asked to elaborate on the example of ADM that they were considering.

Approximately 43% of respondents indicated that they were thinking of ADM in general, or did not provide a specific example. Among the 548 respondents who did provide an example, the most frequent topics/areas mentioned were:

- **Health** (e.g. decisions concerning health insurance, diagnosis of diseases, suggestions for treatment);
- **Justice** (e.g. automation of the judicial system, decisions concerning sentences, fines or early parole);
- **Commerce** (e.g. credit rating, offers for mortgages or loans, customer service, product recommendations);
- **Media & Politics** (e.g. influence in elections, voting advice, news recommendations, advertisement);
- **Self-Driving Cars & Traffic** (including automated control of roads);
- **Work** (e.g. automation of work, replacement of human labour by robots, planning of work schedules, evaluation of job applications).

A similar trend was found in a prompted question about key sectors for AI. Justice and Health ranked the highest, followed by Safety, Education, Economy, Media and Politics.

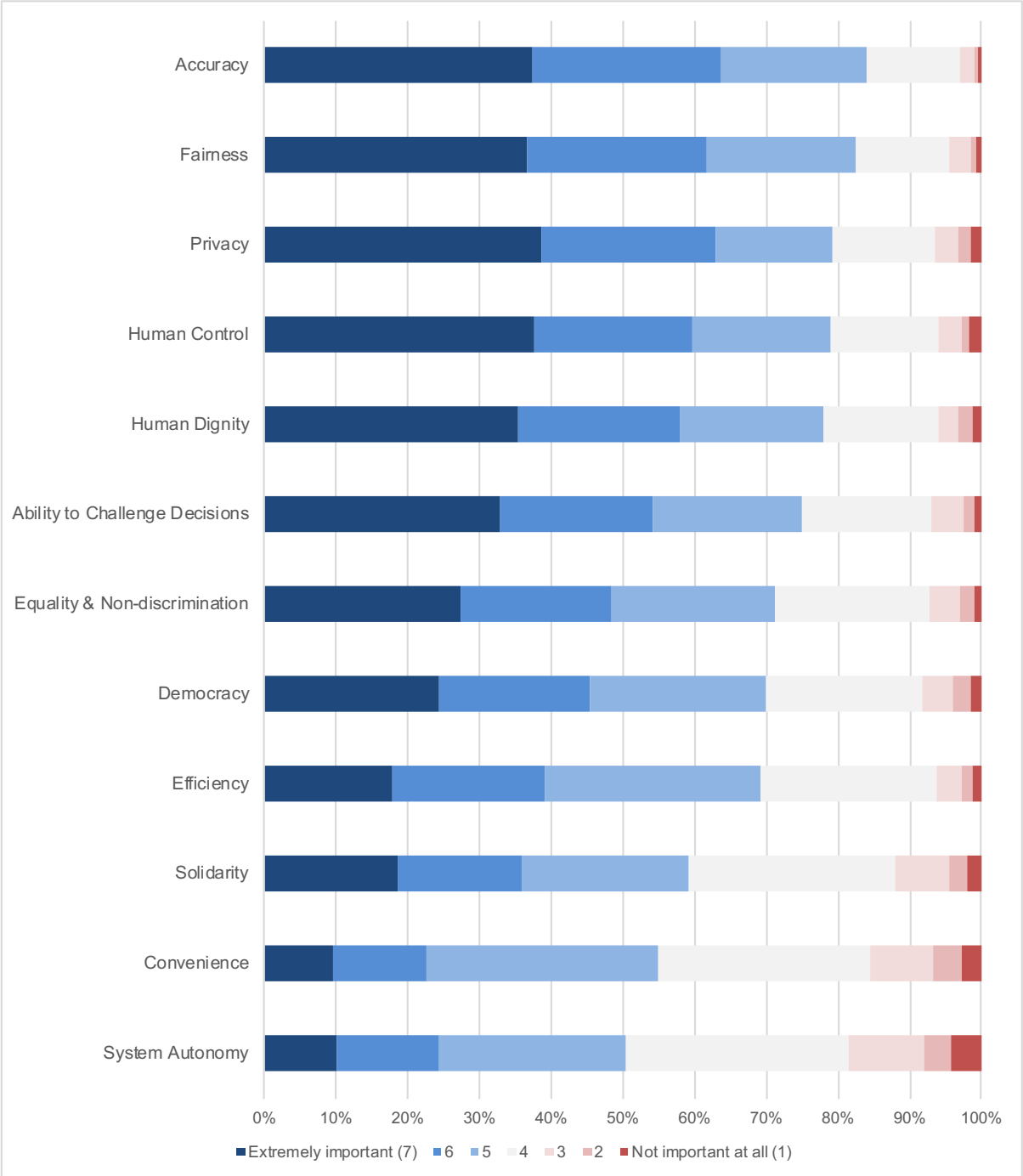
Figure 6. Most mentioned sectors for ADM by AI



Values respondents find important for ADM by AI

Finally, respondents evaluated a set of values for Automated Decision-Making by AI, indicating the extent to which they were important, and prioritising the most important values that should be considered when ADM is in place.

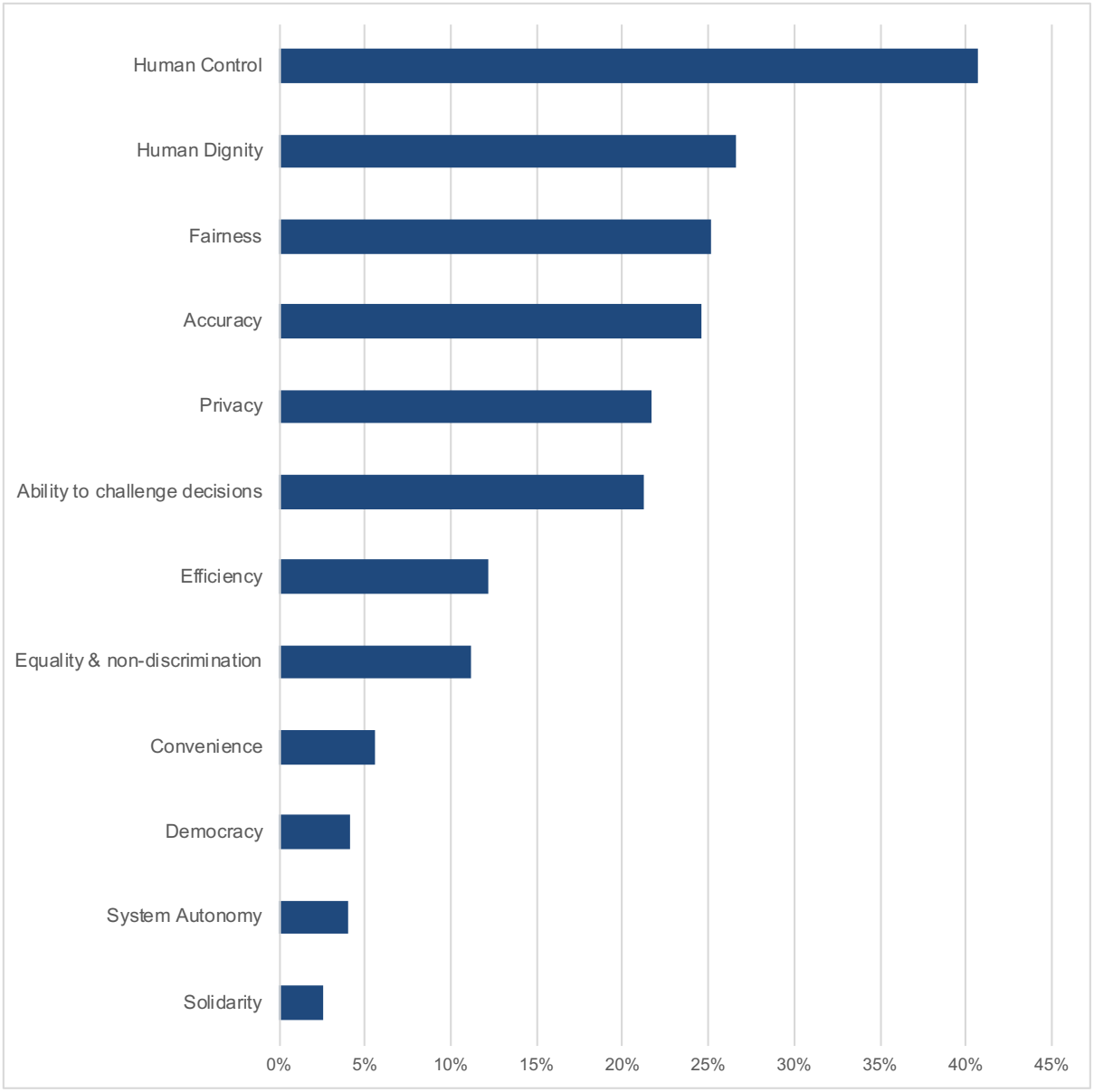
Figure 7. Values for ADM by AI: Importance



In terms of overall importance, Accuracy (*M* = 5.81, *SD* = 1.20), Fairness (*M* = 5.74, *SD* = 1.26), Privacy (*M* = 5.69, *SD* = 1.41) and Human Control (*M* = 5.66, *SD* = 1.39) had the highest average levels of importance.

Respondents were also asked to choose the two most important values for ADM by AI from a list of a possible twelve. When asked to prioritise these values, Human Control (selected by about 41% of the respondents), Human Dignity (27%), Fairness (25%) and Accuracy (25%) were the most chosen.

Figure 8. Prioritised values for ADM by AI



This report is part of a research collaboration between the University of Amsterdam, Tilburg University, Eindhoven University of Technology, Utrecht University and Radboud University on automated decision-making, and forms input to the groups' research on fairness in automated decision-making. Technical details about the report are available in the following pages.

About this report

The questionnaire was developed by a consortium of researchers at the University of Amsterdam, Utrecht University, Radboud University, Tilburg University, and Eindhoven University of Technology. It was sent to respondents provided by Kantar Public, an ISO certified panel, between 26 June and 4 July 2018. The questionnaire used CAWI (Computer assisted web interviewing)[3].

This research was sampled among the members of the NIPObase database, with over 115,000 respondents[4]. A random sample of 3,072 people were invited, out of which 1,069 started filling out the questionnaire. A total of 958 respondents provided informed consent for participation and completed the full questionnaire. The 111 people who did not provide their informed consent did not answer any additional questions. The total response rate among the people invited was 31.2%. This questionnaire was part of a larger project by the consortium.

Results included in this report are unweighted and consider the full sample of valid responses (N = 958). Respondents were required to answer all questions; as such, there are no missing items in data, except for questions following open-ended questions in which "I don't know" was offered as an option (e.g., level of confidence on explanation about AI). Figure 4 (Key sectors for AI) reports the outcome of a content analysis of open-ended responses performed at the University of Amsterdam.

[3] Device type was registered: mobile (26.9%), tablet (2.2%), desktop (12.5%) and large-desktop (57.3%) with 1% of the responses not having a device detected. The questionnaire was compatible with all devices, adjusting for screen size automatically.

[4] Respondents were drawn from NIPObase, Kantar Public's panel with more than 115,000 respondents. The gross sample was representative of the Dutch 18 years+ population on gender, age, education level and region, with random sampling applied within each stratum. Sample was on individual level.

Sample

The composition of the sample was as follows:

	Started questionnaire (N = 1069)	Completed questionnaire (N = 958)	Dutch adult population[5]
Gender			
Male	50.5	51.0	49.2
Female	49.5	49.0	50.8
Highest level of education			
Geen onderwijs\Basisonderwijs	4.3	3.4	5.0
LBO \ VBO \ VMBO (kader- en beroepsgerichte leerweg) \ MBO 1	12.5	11.1	13.3
MAVO \ eerste 3 jaar HAVO en VWO \ VMBO (theoretische en gemengde leerweg)	4.8	4.4	5.0
MBO 2, 3, 4 of MBO oude structuur	33.9	34.0	36.0
HAVO en VWO bovenbouw	5.1	4.6	4.3
HBO-\WO-propedeuse \ HBO-\WO-bachelor of kandidaats	25.3	27.0	24.2
HBO-\WO-master of doctoraal	14.2	15.4	12.2
Region			
Drie grote gemeenten (Amsterdam, Rotterdam, Den Haag)	11.7	11.5	11.8
West (Utrecht, Noord-Holland, Zuid-Holland excl. drie grote gemeenten en randgemeenten)	30.6	31.0	29.3
North (Groningen, Friesland, Drenthe)	11.4	11.3	10.1
East (Overijssel, Gelderland, Flevoland)	21.5	21.0	20.8
South (Zeeland, Noord-Brabant, Limburg)	20.6	20.9	24.0
Randgemeenten (Amstelveen, Diemen, Landsmeer, Ouder-Amstel, Ridderkerk, Barendrecht, Albrandwaard, Krimpen a\d IJssel, etc.)	4.2	4.4	4.0
Age group			
18 - 19	2.1	2.0	3.0
20 - 24	4.5	4.1	7.9

[5] Source: Gouden standaard for representative samples in the Netherlands provided by MOA (MOA, Expertise Center voor Marketing-insights, Onderzoek & Analytics) as provided by Kantar Public.

25 - 29	6.8	7.0	7.9
30 - 34	6.4	7.0	7.5
35 - 39	7.7	8.2	7.5
40 - 44	7.9	8.2	8.3
45 - 49	8.2	8.6	9.6
50 - 54	10.8	10.9	9.5
55 - 59	10.1	10.2	8.7
60 - 64	9.0	9.0	7.8
65 - 69	10.1	9.9	7.8
70 or older	16.6	14.9	14.4
Note: items above may not add up to 100 because of rounding.			

Overview of items included in this report

Explanation about ADM by AI

Om te beginnen zouden we graag een beeld willen krijgen van de mate waarin u, in het algemeen, bewust bent van **kunstmatige intelligentie of van computers** die **beslissingen** nemen op basis van data **zonder de betrokkenheid van mensen**.

Zou u in een aantal woorden kunnen uitleggen wat u verstaat onder **automatische besluitvorming door kunstmatige intelligentie of door computers**?

Confidence in explanation provided by the respondent

Hoe zeker bent u ervan dat uw uitleg van "**automatische besluitvorming door kunstmatige intelligentie of computers**" juist is?

Self-reported knowledge

Hoeveel kennis heeft u van ...?

algoritmen

kunstmatige intelligentie

Hopes and Concerns

Geautomatiseerde besluitvorming door kunstmatige intelligentie of door computers kan worden gedefinieerd als computerprogramma's die beslissingen nemen die voorheen door mensen werden genomen. Deze beslissingen worden automatisch door computers genomen op basis van data.

We willen u hier graag een aantal vragen over stellen.

Als geautomatiseerde besluitvorming meer gaat voorkomen, in hoeverre bent u het dan eens of niet eens met de volgende stellingen wanneer het gaat om **de gehele samenleving**?

Het zal nuttig zijn
Het zal het leven verbeteren
Het zal leiden tot eerlijkere beslissingen
Het zal leiden tot objectieve behandeling

Het zal risicovol zijn
Het kan tot onaanvaardbare resultaten leiden
Het zal leiden tot bezorgdheid
Het kan leiden tot manipulatie

Examples – used for categorization of sectors associated with ADM by AI

Zou u een voorbeeld kunnen geven van geautomatiseerde besluitvorming waar u aan dacht bij het beantwoorden van de voorgaande vragen?

Perceived importance

Wanneer u denkt aan de belangrijkste trends in de nabije toekomst (5 tot 10 jaar), kunt u dan aangeven hoe belangrijk de volgende onderwerpen zijn **voor de gehele samenleving**?

Computers of kunstmatige intelligente ter vervanging van menselijke besluitvormers

Gezondheids- en ouderenzorg

Criminaliteit en veiligheid

Samenleven / normen en waarden

Natuur en milieu

Sociaal stelsel, verzorgingsstaat

Werkgelegenheid

Values for ADM

Wanneer u denkt aan **geautomatiseerde besluitvorming door kunstmatige intelligentie of computers**, in welke mate zijn onderstaande **waarden** dan belangrijk bij het maken van deze systemen die, namens ons, automatisch beslissingen nemen?

Gemak

Eerlijkheid

Privacy

Nauwkeurigheid

Solidariteit

Menselijke waardigheid

Vermogen van het systeem om zelfstandig te werken

Menselijke controle

Efficiëntie

Gelijkheid en non-discriminatie

Vermogen om de beslissing te bevechten

Democratie

Prioritisation of values

Wat zijn voor u de twee belangrijkste waarden voor geautomatiseerde besluitvorming door kunstmatige intelligentie of computers uit de onderstaande lijst?

(Same values shown as above)

References

1. Nationale wetenschapsagenda. (2015). *Dutch National Research Agenda: questions, connections, prospects* (pp. 1–216). Retrieved from https://wetenschapsagenda.nl/publicatie/nationale-wetenschapsagenda-nederlands/
2. VSNU. (2017). *Digital Society Research Agenda: leading the way through cooperation in a digital society* (pp. 1–38). Retrieved from https://www.thedigitalsociety.info/digital-society-research-agenda/
3. Ministerie van Economische Zaken. (2016). *Digitale Agenda: Vernieuwen, vertrouwen, versnellen* (pp. 1–43). Retrieved from https://www.rijksoverheid.nl/onderwerpen/ict/documenten/rapporten/2016/07/05/digitale-agenda-vernieuwen-vertrouwen-versnellen
4. European Commission. (2018). *Artificial intelligence: Commission outlines a European approach to boost investment and set ethical guidelines* (No. IP/18/3362) (pp. 1–2). Retrieved from http://europa.eu/rapid/press-release_IP-18-3362_en.htm
5. Muñoz, C., Smith, M., & Patil, D. (2016). *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights* (pp. 1–29). Executive Office of the President (United States). Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf
6. Larus, J., & Hankin, C. (2018). Regulating automated decision making. *Communications of the ACM*, *61*(8), 5–5. doi:10.1145/3231715
7. Dekker, P., van der Ham, L., & Wennekers, A. (2018). *Burgerperspectieven* (No. 2018|1) (pp. 1–56). Sociaal en Cultureel Planbureau. Retrieved from

https://www.scp.nl/Publicaties/Alle_publicaties/Publicaties_2018/Burgerperspectieven_2018_1