Scriptie: “Vertrouwen in en acceptatie van hoogrisicobeslissingen genomen door mensen en algoritmen”

Introductie

Deze scriptie onderzocht het verschil in vertrouwen in en de acceptatie van hoogrisicobesluiten wanneer deze door mensen of algoritmen zijn genomen in drie verschillende scenario’s. Het onderzoek richtte zich op algoritmesubjecten – mensen over wie een algoritmisch besluit wordt genomen – en hun reactie op de aard van de besluitvorming.

De aanleiding is de toenemende inzet van algoritmen door organisaties om automatisch besluiten te laten, evenals de Algemene Verordening Gegevensbescherming en de aankomende Artificial Intelligence Act die transparantie vereisen over het gebruik van (hoog-risico) algoritmen die geautomatiseerde beslissingen nemen, en de mogelijkheid om als betrokkene een menselijke blik te vragen bij die besluiten.

Aan het onderzoek namen 579 inwoners van Nederland tussen de 18 en 65 jaar deel, proportioneel gestratificeerd op basis van geslacht, leeftijd en opleidingsniveau. De deelnemers werden onderverdeeld in vier groepen: een controlegroep die geen uitleg kreeg over wie of wat het besluit had genomen, en drie behandelgroepen die geconfronteerd werden met beslissingen genomen door mensen, algoritmen, of algoritmen waarbij ook een optie voor menselijke herbeoordeling werd genoemd. Het onderzoek maakte gebruik van drie hoog-risico-scenario's: een afwijzing van een sollicitatie, een verkeersboete, en een medisch behandelvoorschrift.

Resultaten

De Spearman-correlaties laten in alle drie de scenario’s een statistisch significant verband zien tussen vertrouwen en acceptatie (ρ = .62, .73 en .83, p < .01). Mensen met meer vertrouwen in de beslissing zullen deze dus eerder accepteren. Met Kruskal-Wallis en post-hoctoetsen is verder het volgende vastgesteld:

- Transparantie over het feit dat een mens de beslissing nam, veranderde het vertrouwen in of de acceptatie van beslissingen niet significant.
- Transparantie over het feit dat een algoritme de beslissing nam, leidde tot significant minder vertrouwen in en acceptatie van het besluit in de scenario's van sollicitatieafwijzing en medisch voorschrift. Dit effect was niet waarneembaar in het scenario van de verkeersboete. Dit suggereert dat de aard van de beslissing (subjectief versus objectief), of de mate waarin deelnemers zich konden inleven dat een besluit geautomatiseerd werd genomen (die hoger was bij het scenario van de verkeersboete, dan de twee andere onderzochte scenario’s) een rol kan spelen op acceptatie en vertrouwen.
- Het aanbieden van een menselijke herbeoordeling had in geen van de scenario’s een significant effect op acceptatie van en vertrouwen in algoritmisch genomen besluiten.

Hoewel algoritmen in bepaalde situaties beter kunnen presteren dan mensen, moet de afname in vertrouwen en acceptatie door algoritmesubjecten een belangrijke overweging zijn bij de inzet van algoritmen in geautomatiseerde besluitvorming. Verder onderzoek is nodig om beter te begrijpen welke factoren bepalen in welke scenario's de acceptatie van en het vertrouwen in de inzet van algoritmen afnemen.

Thesis: 'Trust and acceptance of high-risk decisions made by humans and algorithms'

Introduction

This thesis investigated the difference in trust in and acceptance of high-risk decisions when made by humans or algorithms in three different scenarios. The study focused on algorithm subjects - people about whom an algorithmic decision is made - and their reaction to the nature of the decision-making.

It was prompted by the increasing deployment of algorithms by organisations to make automated decisions, as well as the General Data Protection Regulation and the upcoming Artificial Intelligence Act requiring transparency on the use of (high-risk) algorithms that make automated decisions, and the possibility, as a data subject, to ask for a human perspective on those decisions.

The study involved 579 Dutch residents between 18 and 65, proportionally stratified by gender, age and education level. Participants were divided into four groups: a control group that received no explanation of who or what had made the decision and three treatment groups that faced decisions made by humans, algorithms, or algorithms where an option for human reassessment was also mentioned. The study used three high-risk scenarios: a job application rejection, a traffic fine, and a medical treatment prescription.

Results

The Spearman correlations showed a statistically significant relationship between trust and acceptance in all three scenarios (ρ = .62, .73 and .83, p < .01). Thus, people more confident in the decision are more likely to accept it. Kruskal-Wallis and post hoc tests further established the following:

- Transparency about the fact that a person made the decision did not significantly change confidence in or acceptance of decisions.
- Transparency about the fact that an algorithm made the decision led to significantly lower trust in and acceptance of the decision in the job application rejection and medical prescription scenarios. This effect was not observable in the traffic fine scenario. This suggests that the nature of the decision (subjective versus objective), or the extent to which participants could empathise that a decision was automated (which was higher in the traffic fine scenario than the other two scenarios studied), may play a role in acceptance and trust.
- Offering a human reassessment had no significant effect on acceptance of and trust in algorithmically made decisions in any of the scenarios.

Although algorithms can outperform humans in certain situations, the decrease in trust and acceptance by algorithm subjects should be an important consideration when using algorithms in automated decision-making. Further research is needed to understand better what factors determine in which scenarios acceptance of and trust in algorithm deployment declines.