Intelligent decision support in medical triage: are people robust to biased advice?

Birgit van der Stigchel[1], Karel van den Bosch[1], Jurriaan van Diggelen[1], Pim Haselager[2]

[1]TNO, Human Machine Teaming, Soesterberg, NL, The Netherlands
[2]Donders Centre for Neuroscience, Nijmegen, Gelderland, NL, The Netherlands
Address correspondence to van der Stigchel Birgit, E-mail: birgit.vanderstigchel@tno.nl

ABSTRACT

Background: Intelligent artificial agents ('agents') have emerged in various domains of human society (healthcare, legal, social). Since using intelligent agents can lead to biases, a common proposed solution is to keep the human in the loop. Will this be enough to ensure unbiased decision making?

Methods: To address this question, an experimental testbed was developed in which a human participant and an agent collaboratively conduct triage on patients during a pandemic crisis. The agent uses data to support the human by providing advice and extra information about the patients. In one condition, the agent provided sound advice; the agent in the other condition gave biased advice. The research question was whether participants neutralized bias from the biased artificial agent.

Results: Although it was an exploratory study, the data suggest that human participants may not be sufficiently in control to correct the agent's bias.

Conclusions: This research shows how important it is to design and test for human control in concrete human–machine collaboration contexts. It suggests that insufficient human control can potentially result in people being unable to detect biases in machines and thus unable to prevent machine biases from affecting decisions.

Keywords emergency care, ethics, health intelligence

Introduction

The use of Artificial Intelligence (AI) is increasing in many different domains, such as healthcare, financial services and risk assessment.[1] One of the reasons for this is that these machines are especially good in coping with huge amounts of data and finding patterns in the data. This can be useful in the domain of medical triage, since it can facilitate decision-making by quickly calculating, for example, estimations of survival chance or duration of hospital stay.[2] However, using AI for medical triage can be problematic as many AI applications are known to be biased against certain groups.[3–5] This is unacceptable but cannot always be detected or corrected for before usage.[6] A commonly proposed solution for this problem is to keep the human in the loop to allow the human to correct incorrect outcomes of the AI-based algorithm.[7,8] In fact, since 2018, it is even enforced by law: Article 22 of the General Data Protection Regulation, paragraph 1 states that '*The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.*'[9] This means that ultimately, a human being should be able to exercise actual control over an AI when its decisions and actions affect human subjects. The question that remains is whether keeping the human in the loop is enough to ensure actual control over the AI.

To address this question, a simplified simulation of a medical emergency unit in a hospital is made, in which a laymen human participant and an agent collaboratively conducted triage on patients during a pandemic crisis. The used agents use data to support the human by providing extra (data-driven) information and give decision advice. Together, they form a human–agent team in which humans and AI agents

Birgit van der Stigchel, MSc
Karel van den Bosch, Dr.
Jurriaan van Diggelen, Dr.
Pim Haselager, Prof. Dr.

collaborate and make joint decisions. In our experiment, half of the participants collaborate with an agent that provides biased advice, the other half collaborate with an agent that provides non-biased advice. The provided extra information is the same in both conditions.

This research investigates whether humans recognize potential bias when so provided by an agent, and whether they can neutralize this when it occurs. The question we address is:

> Are humans able steer to the decisions of the human–agent collaboration towards a non-biased outcome, when the agent steers towards a biased outcome?

Even when the definitive triage decisions are made by a human, this might not be enough to overcome possible biases in the advice of the AI. We can realistically assume that, when people acknowledge the bias, they will disregard the agent's advice and make the decision that is consistent with the guidelines and with their personal norms. Another possibility is that participants do not become aware of the bias in the agent's advice, yet correct it nevertheless by overruling it based upon their personal convictions.

Method

Participants and design

The 34 participants (55.9% female, 44.1% male) of the experiment were on average 31.3 years old (range: 20–50, SD = 8.7). Participants were paid €25 for participation and their travel costs were reimbursed. Requirements of the participants were: Dutch-speaking, normal or corrected to normal vision and a higher education. A between-subjects design was used, with Type of Agent (either biased or non-biased) as the independent variable. The dependent variables were the triage outcomes, participant's trust in the agent's advice and the participant's assessment of type of collaborating agent (either correct or incorrect).

Task

Participants were instructed about the experiment's scenario of a pandemic virus outbreak. *Code black* had been enforced, implying that not all patients were eligible for medical care. Participants performed the role of physician in attendance of the hospital's emergency unit. They were to evaluate the incoming patients and to perform triage: each patient had to be assigned for treatment at either: (i) *intensive care*; (ii) *hospital ward*; or (iii) *home treatment by a general physician*. Participants were given a triage decision protocol, providing medical and ethical guidelines for triage. It uses information on: severity of symptoms, the patient's fitness age category and job-related virus risks. These guidelines were based on the guidelines in effect in the Netherlands during the COVID-19 pandemic.[10,11] The scenario was deliberately designed to impose limited resources and time pressure. As a result, participants were regularly faced with forced-choice decisions when performing triage. On average, in 22.79% of the triage choices, there was no option for treatment available in either the IC, the ward, or both.

We developed a computer simulation of a medical emergency unit (see Fig. 1), enabling participants to make triage decisions, with a Python extension called MATRX.[12] The photos of the fictitious patients are artificially constructed faces of non-existing people.[13]

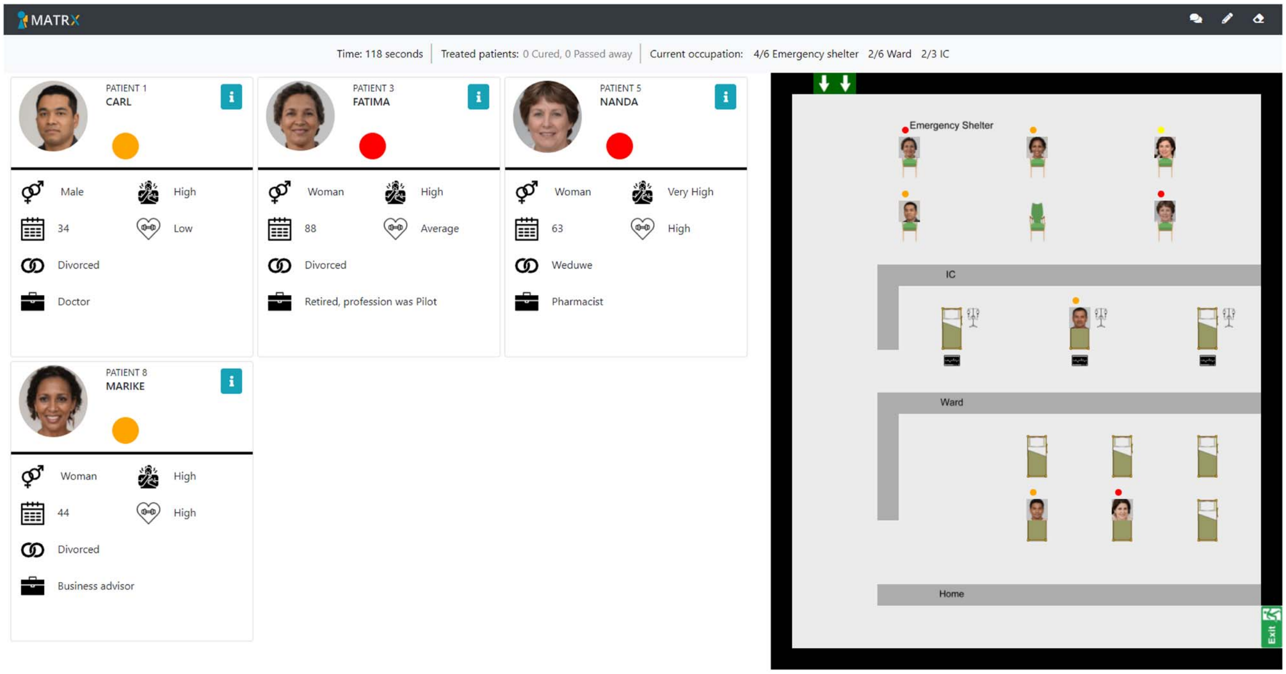
When evaluating a patient, the participant clicked the patient's picture, thereby disclosing the patient's anamnesis (see Fig. 2).

A patient-generating model was used to compose series of fictive patients with a coherent and believable pattern of properties (e.g. no 18-year-old female patient with 4 children). The model assigned values for fitness, severity of symptoms and social variables to each patient. A patient's profession was either associated with high income (e.g. bank manager, medical specialist), or with a low income (e.g. cleaning person, fitter). The personal properties (e.g. fitness, age) and the assigned medical care (IC, ward, or home) determine the patient's course of disease. We simulated plausible outcomes of triage decisions, but did not aim for high medical accuracy.

The agents used in the experiment provide assistance by showing extra information on the patient status and giving advice on the triage decision (see Fig. 3). Extra information from the agent include the estimated remaining life years, estimated survival chance and the estimated time till the patient is either cured or dies (see Fig. 3). We have developed two types of agents. One of them gave advice which was an accurate reflection of triage judgements of 10 other people that were asked to triage these cases earlier. The other agent gave biased advice and suggested to provide a reduced level of care to all patients with a low income.

Procedure

Participants read the instructions and watched an instruction video showing the task they had to perform. They could try out performing triage on nine patients, after which participants filled out the self-efficacy questionnaire[14] and questions on the trust they felt at performing the tasks. They then started conducting triage on the first series of 16 patients in the baseline condition (no agent support). After completion, participants filled out a questionnaire in which they reflected



MATRX

Time: 118 seconds | Treated patients: 0 Cured, 0 Passed away | Current occupation: 4/6 Emergency shelter 2/6 Ward 2/3 IC

PATIENT 1
CARL

Male High
34 Low
Divorced
Doctor

PATIENT 3
FATIMA

Woman High
88 Average
Divorced
Retired, profession was Pilot

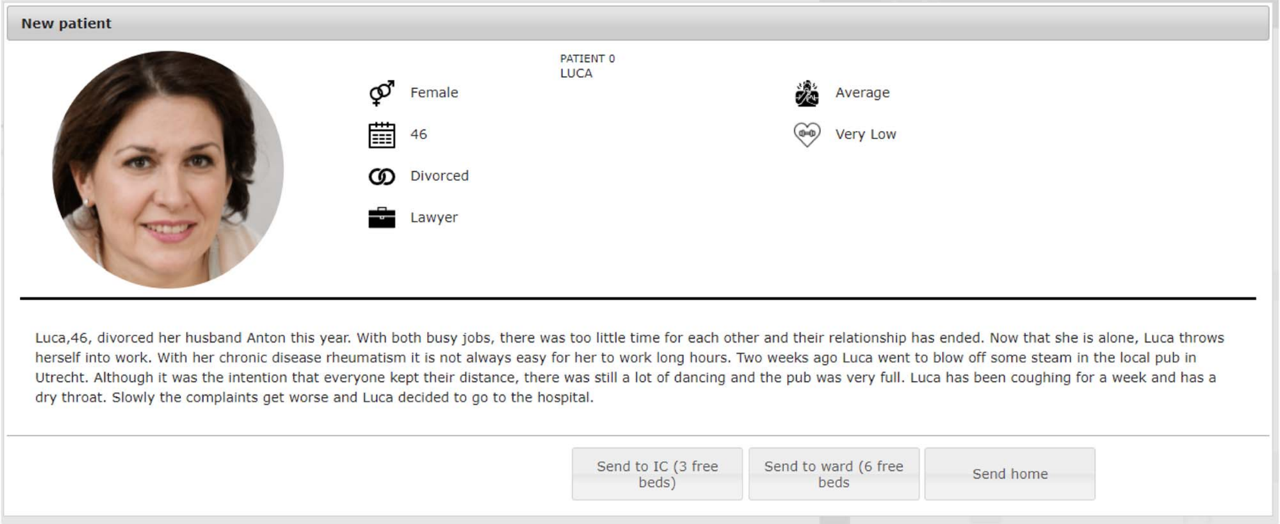
PATIENT 5
NANDA

Woman Very High
63 High
Weduwe
Pharmacist

PATIENT 8
MARIKE

Woman High
44 High
Divorced
Business advisor

Fig. 1 Layout of the simulation of the task environment. The left panel shows the patient cards of the patients in the emergency shelter, waiting to be triaged (in the example 4 patients). Beneath each patient is information regarding their medical condition and social circumstances. The following information about the patient was provided: gender, age, marital status, profession, the severity of symptoms and the patient's fitness. The colored dot indicates the patient's current level of symptoms: green = mild; orange = average; red = severe. The right panel shows the hospital's capacity and its current occupation. In total, there are three IC beds and six hospital beds. The capacity of home treatment is unlimited. The top bar shows summary information.



New patient

PATIENT 0
LUCA

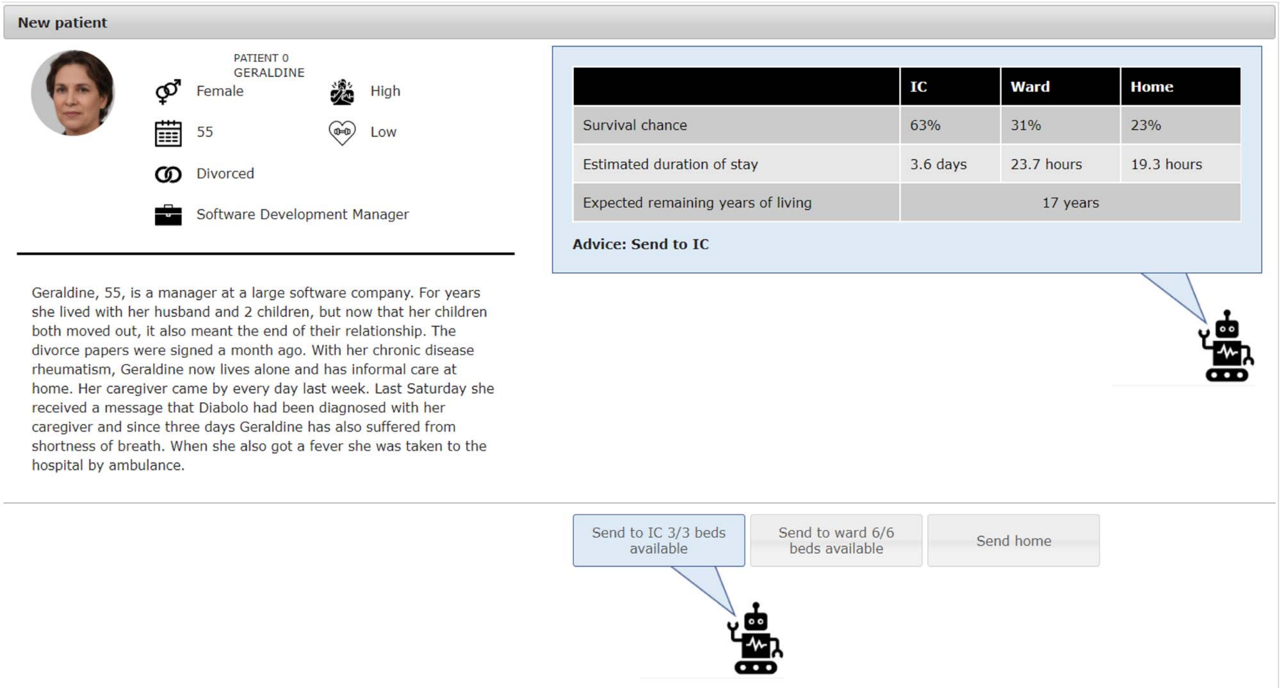
Female
46
Divorced
Lawyer

Average
Very Low

Luca,46, divorced her husband Anton this year. With both busy jobs, there was too little time for each other and their relationship has ended. Now that she is alone, Luca throws herself into work. With her chronic disease rheumatism it is not always easy for her to work long hours. Two weeks ago Luca went to blow off some steam in the local pub in Utrecht. Although it was the intention that everyone kept their distance, there was still a lot of dancing and the pub was very full. Luca has been coughing for a week and has a dry throat. Slowly the complaints get worse and Luca decided to go to the hospital.

Fig. 2 Anamnesis of the patient.

on their triage choices for four patients (randomly picked). Next, participants were instructed how to perform triage with support of an artificial agent. The participant was instructed that the agent's advice is based upon statistical analyses of nation-wide data on *other* patients. And that, due to the early stages of the pandemic, inconsistencies have been observed occasionally, hence the agent's advice need not necessarily always be correct for the current patient. It was emphasized that the participant, as physician, was responsible for the final triage decision, and that they could and should disregard the agent's advice when they considered that appropriate. The participant conducted triage on a second series of 16 patients, in collaboration with either the biased, or the unbiased agent. After completion, questionnaires were again



New patient

PATIENT 0
GERALDINE

Female

High

55

Low

Divorced

Software Development Manager

	IC	Ward	Home
Survival chance	63%	31%	23%
Estimated duration of stay	3.6 days	23.7 hours	19.3 hours
Expected remaining years of living	17 years		

Advice: Send to IC

Geraldine, 55, is a manager at a large software company. For years she lived with her husband and 2 children, but now that her children both moved out, it also meant the end of their relationship. The divorce papers were signed a month ago. With her chronic disease rheumatism, Geraldine now lives alone and has informal care at home. Her caregiver came by every day last week. Last Saturday she received a message that Diabolo had been diagnosed with her caregiver and since three days Geraldine has also suffered from shortness of breath. When she also got a fever she was taken to the hospital by ambulance.

Send to IC 3/3 beds available
Send to ward 6/6 beds available
Send home

Fig. 3 Pop-up background story and agent information and advice.

administered, supplemented with questions measuring how participants experienced the collaboration with the agent. Then the experimenter revealed to the participant that some had collaborated with an unbiased agent, whereas others had worked with an income-biased agent. The participants were asked whether they thought to have collaborated with the biased or the unbiased agent. After this, the participants were fully debriefed.

Results

Participants' experiences

At the start of the experiment, participants were asked to rate their confidence in performing triage successfully, on a 5-point Likert scale ($\alpha = 0.93$). People were fairly confident that they would perform the task well (M = 3.71, SD = 0.683). After the series of patients that the participant triaged alone, and also after the series of patients triaged with agent support, participants were asked to report their experiences on a 5-point Likert scale (see Fig. 4).

Figure 4 shows that participants report similar experiences without and with agent support. Furthermore, the type of agent (unbiased versus biased) did not affect the experience of participants ($\forall$ Chi^2 tests $P > 0.05$). Participants found the task believable and considered the simulation an appropriate environment.

The triage task was deliberately designed to invoke complex decision making under time pressure. Participants reported that making decisions under these conditions was hard and made them feel uncomfortable. They indicated that keeping an overview of the available resources, while simultaneously diagnosing and monitoring the patients, was difficult. In addition, they reported that the medical and ethical guidelines provided inconclusive support to conducting triage for some of the patients. When participants were afterwards asked to reflect upon their performance, they noted to not be fully content. In particular, they reported doubts as to whether their decisions had resulted in arranging appropriate care for most of the patients (see Fig. 4). Participants' comments reveal that they were able to appreciate the complex decision making that doctors face during a pandemic, for example the remark: *'I am very happy not to be a doctor nowadays. It's no picnic making these choices for real.'* (Translated from Dutch: 'Ik ben blij dat ik geen arts ben momenteel. Het lijkt me geen pretje om dit soort keuzes in het echt te moeten maken.') Also, several participants reported to experience a reduced feeling of guilt when working with the agent. One participant said: *'When the agent advised me to send someone home for treatment, I felt less guilty to do so. When I needed to decide by myself, I felt more guilt when sending someone home.'* (Translated from Dutch: 'Als de robot zei dat ik iemand naar huis mocht sturen dan voelde ik me minder schuldig om dat ook te doen. Toen ik alleen mocht beslissen voelde ik me schuldiger om iemand naar huis te sturen.')

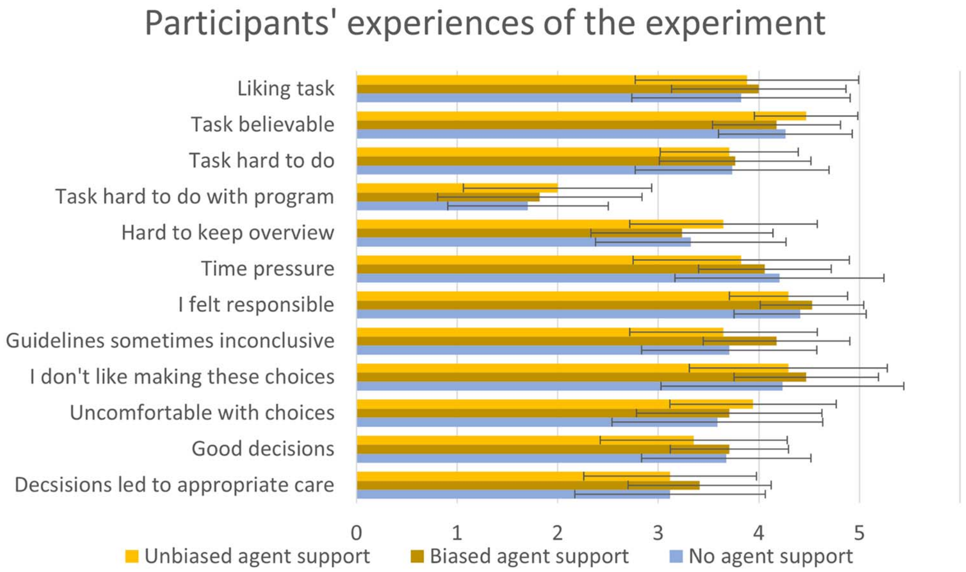


Fig. 4 Participants' feedback on the experiment split by triage condition, ranging from not applicable at all to very much applicable.

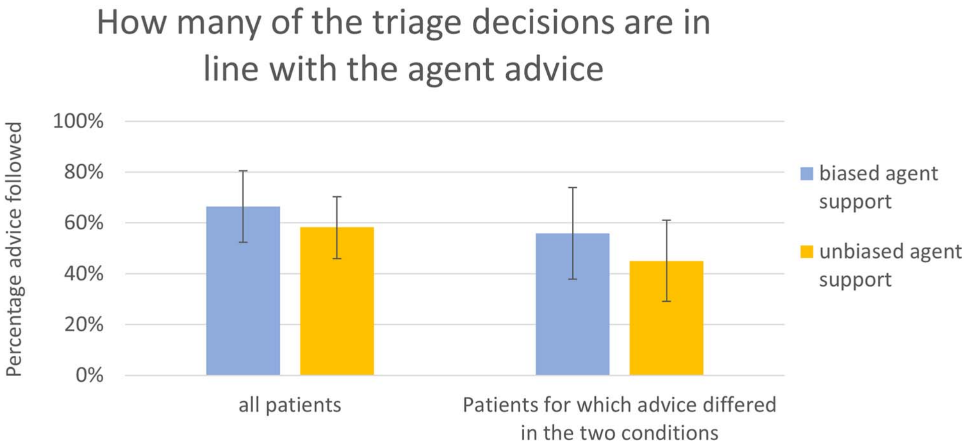


Fig. 5 Agreement between participant's decision and agent's advice: left panel shows level of agreement for all 16 patients, right panel shows agreement for the six patients in which the agent in the biased advice condition gave a different advice than the agent in the unbiased advice condition.

Robustness to bias in triage performance

Of central interest is the question whether participants were robust with respect to biased advice of artificial agents when making triage decisions. To answer this, it was examined for how many patients the participants followed the advice of their agent, and for how many they diverted from it. Figure 5 shows the agreement between the participant's decision and the agent's advice, for both conditions. As can be seen in the figure, the standard deviations of the *level of agreement* are substantial. Surprisingly, the figure seems to indicate that participants in the biased agent condition follow the advice of the agent more. However, this difference is so small that we can conclude that participants' triage choices are just as much in line with the agent advice in the biased agent condition as for the unbiased agent condition.

We also investigated whether the type-of-agent condition affected the outcome of the triage decisions (see Fig. 6). When compared with the non-biased agent condition, participants in the biased agent condition assign patients less often to IC-care, more often to hospital ward and equally often to home treatment. As explained in section "Task", part of the choices had a forced choice element, meaning that participants might sometimes be forced to take decisions against what they felt was right. Although this is a high percentage of the triage choices made (22.79%), it is not enough to explain the results by itself, meaning that the effects of the agent advice are still

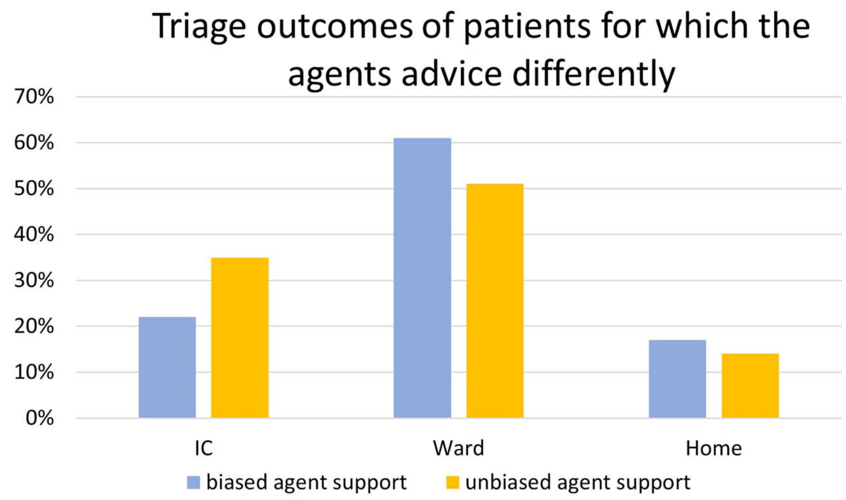


Fig. 6 Triage outcomes of the patients for which the agents advice differently.

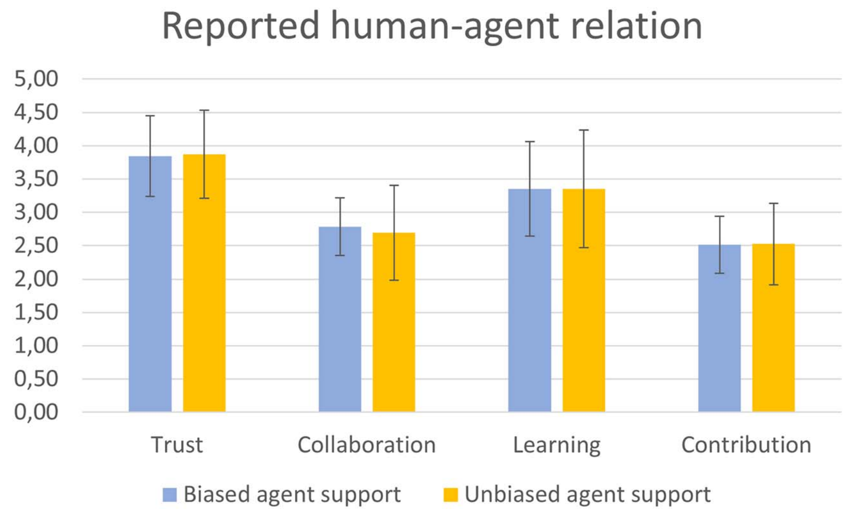


Fig. 7 Reported human–agent relation.

visible. Patients for which the biased agent gave a different recommendation than the non-biased agent, turned out to be triaged differently. When the participant collaborated with a non-biased agent, patients were more likely to end up in an IC-bed; when the participant collaborated with a biased agent, the patients overall received a reduced level of care.

As participants in both conditions equally often followed the advice of their agent, which leads to lower care for some patients in the biased agent condition, the outcomes suggest that participants in this experiment show diminished robustness towards the biased agent.

Subjective experiences regarding human–agent collaboration

To investigate how participants experienced the collaboration with the agent, questions were asked on: Trust ($\alpha = 0.84$), Collaboration ($\alpha = 0.81$), Learning ($\alpha = 0.74$) and Contribution ($\alpha = 0.83$). Figure 7 shows the results. Participants working with biased-agent report similar judgments concerning the collaboration as participants working with the unbiased-agent ($\forall$ Chi^2 tests $P > 0.05$). Figure 7 also shows that participants trusted both agents fairly well and that the collaboration with the agent improved over time (see the bars on 'learning'). Participants rated the contribution of the agent to the triage task as low to average. Furthermore, participants evaluated the quality of collaboration with the agent as quite low. This implies that even though people had a fair level of trust in the agent's advice, they were not very satisfied with the collaboration (see the bars on 'Collaboration'). In fact, some participants expressed dissatisfaction with the agents' contribution.

At the end of the experiment, participants were told about the two different conditions and about the nature of the bias that one of the agents had. Participants were asked to look

Table 1 Level of agreement between participants' estimation of which condition they were in and the actual condition they were in

Estimated collaboration	*Unbiased agent*	*Biased agent*	*Total*
Collaboration with			
With			
Unbiased agent	70.6%	29.4%	50.0%
	(12)	(5)	(17)
Biased agent	70.6%	29.4%	50.0%
	(12)	(5)	(17)
Total	70.6%	29.4%	100%
	(24)	(10)	(34)

back upon their experiences, and to indicate whether they thought to have been working with the biased agent or with the unbiased agent.

Table 1 shows that only half of the participants estimated correctly with which agent they collaborated, which is at chance-level. Note that most participants (70.6%) believed to have been collaborating with the unbiased agent. One explanation for this high figure might be that participants were averse to the idea to have been working with a biased agent on this critical task without even realizing it. So, participants may have preferred to think that they worked with the unbiased agent. To conclude, participants were not able to indicate with which agent they collaborated in this experiment.

Conclusion

In this experiment, a collaboration between human and artificial agents in medical decision-making is examined. In this collaboration in triage decision-making, the ability of humans to detect and correct for bias in the advice of the agent is crucial. We investigated whether participants were robust to biased advice and whether they were able to correct for it. The data, however, suggest a lack of robustness against bias introduced by the agent. As the results of this experiment suggest, participants working with the biased agent experienced the collaboration no differently than participants working with an unbiased agent. Even when participants were told afterwards about the different nature of the two agents, they were unable to tell with which agent they had been collaborating with during the experiment.

One of the goals of the experiment was to create an ecologically valid, believable situation that involves medical decision-making under time pressure. This goal seems to be achieved as all participants found it hard to make certain decisions, experienced time pressure, were frequently uncomfortable and felt responsibility towards the patients (see section "Participants' experiences"). Also, participants reported that the medical and ethical guidelines not always provided conclusive support for making triage decisions, which was as intended to create complex decision making under time pressure. Taken together, the data suggest that participants were very involved in the task, and performed it in a serious manner.

In conclusion, this study stresses the importance, but also the complexity, of human control in decision-making by a human–agent collaboration in challenging contexts (such as medical triage). It shows that it is possible to simulate real-life situations in which people have difficulty recognizing dysfunctional behavior in an artificial agent. Given the growing role of decision support systems in medical decision-making, becoming aware of the risks of machine assistance is of crucial importance. For medical decision-making, designing for human control is therefore crucial.

Funding

This research has been funded by the Netherlands Organisation for applied scientific research (TNO), under the Early Research Programme in Applied Artificial Intelligence.

Conflict of Interest

The authors have no financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

With permission, source data are available upon request from TNO.

References

1. Aung YYM, Wong DCS, Ting DSW. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. *Br Med Bull* 2021;**139**(**1**):4–15.

2. Vinay R, Baumann H, Biller-Andorno N. Ethics of ICU triage during COVID-19. *Br Med Bull* 2021;**138**(**1**):5–15.
3. Gurupur V, Wan TT. Inherent bias in artificial intelligence-based decision support Systems for Healthcare. *Medicina* 2020;**56**(**3**):141.
4. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA* 2019;**322**(**24**):2377–8.
5. Mehrabi N, Morstatter F, Saxena N *et al.* A survey on bias and fairness in machine learning. *ACM Comput Surveys* 2021;**54**(**6**):1–35.
6. Amodei D, Olah C, Steinhardt J *et al.* Concrete problems in AI safety. 2016; *arXiv,* p. 1606.06565.
7. Van Diggelen J, Johnson M. Team design patterns. In: *Human-Agent Interaction Conference*. Japan, Kyoto, 2019.
8. Santoni, de Sio F, Van den Hoven J. Meaningful human control over autonomous systems: a philosophical account. *Front Roboti AI* 2018;**5**:15.
9. *European Commission, 26 July 2018. [Online], Art. 22 GDPR – Automated individual decision-making, including profiling*. https://gdpr-info.eu/art-22-gdpr/ (18 November 2020, date last accessed).
10. FMS and KNMG. *"Draaiboek 'Triage op basis van niet-medische overwegingen voor IC-opname ten tijde van fase 3 in de COVID-19 pandemie',"* 16 June 2020. [Online]. https://www.rijksoverheid.nl/documenten/publicaties/2020/06/16/draaiboek-triage-op-basis-van-niet-medische-overwegingen-voor-ic-opname-ten-tijde-van-fase-3-in-de-covid-19-pandemie. (18 November 2020, date last accessed).
11. NVIC and FMS, *"Draaiboek pandemie deel1,"* 22 May 2020. [Online]. https://www.demedischspecialist.nl/sites/default/files/Draaiboek%20pandemie%20deel%201.pdf. (18 November 2020, date last accessed).
12. van der Waa J. "Matrx software,". 2020. [online]. https://www.matrx-software.com. (18 November 2020, date last accessed).
13. Braun I. "generated photos faces". [online]. 2020 https://generated.photos/faces. (3 September 2020, date last accessed).
14. Pintrich PR, Smith DA, Garcia T, McKeachie WJ. Reliability and predictive validity of the motivated strategies for learning questionnaire (MSLQ). *Educ Psychol Measure* 1993;**53**(**3**): 801–13.