Bachelor thesis

Artificial Intelligence

Radboud University

The Functional Preference of Policemen when Using a Reflection Machine in Preventive Frisking

Author:
J.J. Venhuizen
s1016521

First supervisor:
dr. W.F.G. Haselager
Artificial Intelligence
w.haselager@donders.ru.nl

Second supervisor:
dr. H.K. Schraffenberger
Digital Security &
Interdisciplinary Hub for
Digitalization and Society
hanna.schraffenberger@ru.nl

January 28, 2022

Abstract

In this thesis we dive into the concept of a Reflection Machine (RM) and the functional preference of policemen interacting with it in a preventive frisking context. An RM is a system which aims to improve the user's decision strategy and enhance their reflection process by questioning the user's decision. Concretely in this thesis, we try to accomplish this by the use of reflective interventions which can be embodied by either questioning or notifying the user about specific aspects. We implemented a hard-coded version of this conceptual RM in an application simulating a preventive frisking context and used it in an experiment where we questioned the participants about the experience with the RM afterwards. By analyzing this, we show that interventions based on questioning the user's degree of certainty over their decision is experienced as most relevant in general. Furthermore, we argue that we are uncertain about the degree of influence of a conceptual decision support system on the experience of an RM. Finally, we show that the users generally like the interventions and argue that an RM may be a valuable tool in the police domain.

Contents

Chapter 1

Introduction

Observers observing observers has become reality for the situation in preventive frisking performed by policemen in Amsterdam [1]. The concept of preventive frisking entails frisking people who are present in a so-called security risk area assigned by the municipality in collaboration with the public prosecutor and the police [2], [3]. Currently in Amsterdam, a-selective preventive frisking is performed which is realized by randomizing the process into frisking every fifth passenger. Though, in order to make sure the police does this in a fair way, the municipality assigned 45 citizens to observe the actions of policemen when individuals in an area are frisked in a preventive manner. According to the Dutch police association [1], this is experienced as a lack of trust in policemen [4]. They state that they are being accused of bias, mainly racial bias. This accusation motivated the police to assign observers to observe the observers. Fortunately, there are already existing techniques which try to overcome the problem of potential bias in policemen by supporting (self-)reflection in policemen. However, these techniques carry disadvantages with them [2]. Due to the shortcomings we think that it is desirable to implement a new system in this field which overcomes these disadvantages.

A Reflection Machine (RM) could embody this new system. An RM is a system that aims to improve the user's decision strategy and enhance their reflection process by questioning the user's decision [5]. Self-reflection is associated with improving a dynamic decision making process by reducing cognitive biases [6]. Following this line, we expect reflective interventions from an external system may help to increase bias awareness by the user. Furthermore, we argue that it can be used to increase meaningful human control; transparency about decisions to other people; and confidence of the user. In turn, this could result in a re-build of trust in the police such that observers observing observers are not necessary anymore. In order to be

[1]Translated from Dutch: "Nederlandse Politie Bond (NPB)"

[2]See 2.4.2

able to truly support the police, we need to ensure that the RM satisfies the needs of policemen. This is why we aim to research the functional preference of policemen when interacting with a RM in preventive frisking. We do this by making use of an application implemented by Anaïs de Graaf [7]. In this thesis we examine three aspects of the functional preference of an RM. We defined these aspects into sub-questions:

1. **What type of reflective interventions do policemen experience as relevant?** We presume this question is important because a reflective intervention needs to be experienced as relevant in order to have an effect; otherwise, sooner or later, users will likely not interact (appropriately) with the RM anymore.

2. **With what frequency should the reflection machine pose a reflective intervention?** We believe the frequency is of high importance for the experience. One can intuitively imagine that too many interventions might be experienced as annoying or critiquing, whereas too few might have no -or little- effect.

3. **Does the experience of interacting with a decision support system influence the experience of the reflection machine?** Where experience of the RM is measured by the level of quality of the reflective interventions and the level of feeling criticized by the reflective interventions indicated by participants of this study. A decision support system (DSS) is a machine that aims to help people make rational decisions by providing suggestions or recommendations. We included this question because we are curious whether users will experience the reflection machine in a different manner when being (falsely) informed about a decision support system.

In the coming chapter, we will dive deeper into the concept of a general reflection machine and the expected effects of it. Next to that, we will further explain the problem of bias and we argue why an RM could contribute value to the field of policing. In the third chapter, we will focus on studies related to the RM; the name may be new, but a similar concept already exists. In chapter four, we will explain the details of our study design consisting of an expert interview and an experiment including a hard-coded RM. In chapter five we analyze the results from both the interview and the experiment. In the last chapters, we end with the conclusion and discussion.

Chapter 2

Preliminaries

2.1 Reflection Machine

As for this moment, an RM does not exist and is not explicitly known in literature. We define it as a machine which aims to improve the user's decision strategy by enhancing their reflection process. It tries to accomplish this by generating reflective interventions during the decision making process, inspired by the reflection-in-action as originally described by Schön [8]. The underlying assumption for improving the decision strategy is that (externally stimulated) reflection may improve certain aspects like increasing bias awareness, transparency, meaningful human control and confidence. Literature shows that reflection is increasingly used in medical school to raise awareness about and reduce the effects of bias [9]. Next, it is trivial that it increases transparency when used in an honest manner, as explaining yourself implies becoming more transparent. Research supports the phenomenon of increased confidence when reflective practice is applied [10].

2.2 Meaningful Human Control and the responsibility gap

Does research support the idea of increased meaningful human control (MHC) by applying reflective practice? Whereas the concept of MHC is mainly used in the autonomous weapon system discussion, Santoni de Sio and van den Hoven argue that it can play a broader role in the responsibility innovation in robotics and AI [11]. To better understand MHC, we first explore the concept of the responsibility gap.

In several domains, experts already make use of decision support systems (DSS's) e.g. a DSS in a medical setting might output the following sentence 'with these symptoms, there is a chance of 98% that the patient suffers from lunge cancer'. When a physician follows this advice and treats this patient accordingly and it turns out that the advice of the algorithm was misleading (e.g. the fed data includes an over-representation of lunge cancer cases), the question of responsibility attribution arises; by virtue of machine learning, algorithms can change their behaviour over time and thus the programmer cannot be held fully responsible. However one can argue to what extent the physician can be held responsible as he is inclined to trust and follow the advice of the DSS because e.g. 1) in the first hundred cases, he did not rely on this advice, but did notice that every single time the advice turned out to be right; and 2) he may be overwhelmed by the high percentage [1]. This indicates a responsibility gap [12].

In a paper studying the ability of people to withstand biased advice from morally dysfunctional agents in the medical field, they conclude that it is possible to simulate involvement creating situations in which people have difficulty recognizing dysfunctional behavior in an artificial agent. They stress the importance, but also the complexity, of meaningful human control in decision-making by a human-agent team in ethically challenging contexts [13].

We believe it is desirable to keep a human being in the loop and keeping this individual as responsible as possible for the final decision. By enhancing the reflection process of a DSS user, we suppose it is possible to diminish this responsibility gap and increase meaningful human control. We can make it more general by posing it is possible to increase MHC by creating more reason responsiveness and enhancing the reflection process of anyone. Though, this assumption needs to be verified by further research.

2.3 Bias in general

Unfortunately, human beings tend to hold on to their prejudices because of three concepts. 1) Whenever we have a belief about the world, we tend to act out of that belief. In this way, we may unintentionally find ourselves in places where we are externally stimulated with information supporting our belief, e.g. an algorithm used to show you specific YouTube videos you like based on your history. Another example is predictive policing where an algorithm suggests a policeman to go towards a certain location because there is a lot of data of criminality, which increases the amount of data for this area [14]. We call this the *'you find more where you search more'* principle.

2) Furthermore, we experience a so-called confirmation bias. This entails that human beings tend to hang on to favoured hypotheses with unwarranted tenacity and confidence. In this way, we often focus on and store information which supports our expectations and strengthens our beliefs about the world

[1] Note that this example is slightly superficial, as the physician is naturally end responsible here. Though, one can intuitively understand that the experienced level of responsibility is moderately diminished here.

caused by internal selection.

3) On top of that, there are studies which conclude that we experience a bias blind spot as indicated by Pronin, E., Lin, D. Y., & Ross, L. [15]. This bias blind spot is the cognitive bias of recognizing the impact of biases on the decisions of others, while failing to see the impact of biases on one's own decisions.

These three concepts make it difficult to become aware of our own bias, let alone get rid of the influence on our behavior.

2.4 Reflection Machine in Policing

We have argued that a reflection machine could have benefits when used together with for example a DSS. However, in the police domain no such system is used, so why include an RM in the police domain? We will explain this by explaining the role of bias in policing. Then, we show why we believe the existing techniques are not enough to solve this problem, and an external reflection tool is of value.

2.4.1 The Problem of Bias in Policing

As aforementioned, the police experiences the observers assigned by the municipality as an accusation of acting out of bias. Specifically the question whether a racial bias exists in policing in the Netherlands is more complicated than a simple yes or no answer, as indicated by Huisman, de Keijser and Weerman [16]. Yet, they conclude that there is an over-representation of non-Dutch civilians with respect to the population distribution when the police is acting on its own initiative in a proactive manner. Next to a potential racial bias, there are also other forms of biases (e.g. age, gender, way of clothing, etc.), which an individual may be subject to.

There are multiple issues with these biases as explained in section 2.3. Whether biases in the police domain exists or not, nevertheless citizens might still hold the assumption that bias is included in actions. And thus the problem of bias exists either way.

2.4.2 Existing Techniques to Overcome Bias

We will name three existing techniques in policing to overcome bias. First, policemen are usually not operating alone which creates a fruitful environment for peer feedback. Among other things, this is effective for double-checking the colleague's acting, which can diminish the role of bias. Second, in the police academy they work with virtual reality (VR) training in order to simulate several scenes (e.g. a robbery at a train station). In the simulation the trainees receive feedback on their acting, which aims to increase responsibility. Third, policemen use the application MEOS (mobile more effective in the streets) [2] in which they can see and note certain facts about individuals. This can give insight and transparency of the users.

However, there are several disadvantages to these solutions. Peer feedback inevitably includes a social factor. This social factor can e.g. entail social pressure, social hierarchy or even in extreme cases coercion. This makes the peer feedback less reliable and transparent. VR training currently only consists of static simulated environments in which police in training are taught how to handle certain situations. These are also used to get insight in and reflect on certain personal prejudices. However, this is not the main focus when making use of VR training. Next to that, it stops after training as they are expected to do not let go of that reflection when they are actually working. Concerning the MEOS application, it requires an immense amount of work and time to analyze every specific individual causing the feedback to be received after a long time. In this scenario, no real-time reflection is involved and the individual might not remember the particular incident. To overcome these disadvantages, a new system should ideally be without social influence, in real-time, case-bound and transparent. An RM may be able to embody these requirements.

2.4.3 External reflection

In a paper about the police as reflective practitioner, Staller and Koerner (2020) come up with four different lenses that aim to uncover action-guiding assumptions of the learners: perspective of the addressee; lens of scientific knowledge; perspectives from colleagues and; experience as addressees. While they find that perspectives from inside the system (i.e. colleagues and own experiences) were engaged with more easily and adopted less critically, they argue that it is known for learning processes in general that the phase of irritation is a key prerequisite for change at the individual and organizational level. Thus, perspectives from outside the system of policing are of particular importance for the reflective police officer. This can be embodied by an RM. The three concepts of bias explained in the previous section are particularly dangerous when an important aspect of a profession is being as neutral and non-prejudiced as possible, as is the case for the police [17]. Therefore an RM could be a great means to minimize the problem of bias by increasing the bias awareness, transparency, meaningful human control and confidence. Note that measuring the effect of an RM can be precarious in this domain. This is due to the inability to ultimately evaluate the decision made. There is no absolute right or wrong in this domain, whereas in other fields there might be. For example, in the medical field it is often possible to objectively evaluate the judgment of a medical specialist rather than of the police.

[2] Translated from Dutch: Mobiel Effectiever Op Straat

Although the effects may be hard to measure, it does not mean it does not have an effect. Increasing bias awareness can possibly take years to become visible. However, if it does have the desired effect, the effects will be perpetual.

2.5 Preventive Frisking

In this section, we provide some more information on preventive frisking. Preventive frisking can be deployed in a security risk area. This is an area assigned by the municipality in collaboration with the public prosecutor and the police. In the Dutch law, there are two articles reserved for preventive frisking, article 151b and article 174b [2],[3].

Article 151b states that the preventive search instrument is a power of inspection that can be used in the event of a disturbance of public order or in the event of serious fear of a disturbance of public order. This instrument can solely be deployed if an area is stated in writing as a 'security risk area' in a specific time slot by the municipality in collaboration with the public prosecutor. The (assistant) public prosecutor can give a concrete order to search an area. Actions are therefore carefully planned at specific times and within specific areas. This indirect form of selection is allowed, provided there is no selective search within the restricted area.

Article 174b states that in an unforeseen, urgent situation, in consultation with the public prosecutor, the mayor may orally designate an area as a security risk area for a maximum of twelve hours.

In a study, ombudsmen state that searches are not always carried out randomly and that selection is partly left to the common sense, the professionalism or intuition of the police officer [18]. The National Ombudsman is an independent and impartial institution assessing complaints about all aspects of public administration, defends the interests of the citizen and monitors the quality of public services in the Netherlands. They consider this partly selective frisking an insufficient basis for an objective and random application regarding preventive frisking. Selection criteria are also not always established prior to an action. This creates a form of selection that is not transparent and verifiable and creates a risk of unauthorized selection, say the ombudsmen. This calls for a more transparent application of preventive frisking, which an RM could embody.

Chapter 3

Related Work

3.1 Falsification Machine

Although the defined RM is new to literature, there are other theses written about the concept of a Falsification Machine (FM), which embodies a similar system [5]. We used the term RM, because the main goal is reflection and not falsification; where the process of falsifying entails trying to refute a user's decision, reflection tries to make the user more aware of the motivation behind the decision. This is a thin line, but it is important to grasp the difference because the concept of an FM could easily provoke a sense of resistance while the RM could permit a more supporting impression.

3.2 External debiasing technique

The concept of an RM (or an FM) is not entirely new. In 1989, Kydd already writes about the use of an external debiasing technique [19]. He introduces this as a potential solution to the problem rising as a side-effect when human beings use a computer-based DSS. This side-effect difficulty emerges when a user begins to let go of questioning the output of the DSS. They may tend to see it as a truth rather than a supporting tool. This may result in unwittingly eliminating the human out of the decision-making loop. This is why Kydd came up with an external debiasing technique to keep the human inside the loop by trying to falsify the user's end decision.

Chapter 4

Research

4.1 Research setup

In this research we studied the experience of policemen with a Reflection Machine (RM) in preventive frisking. We decided to use a mixed methods setup to study this enabling us to do a combination of qualitative and quantitative research. First, we conducted an expert interview with a policeman in training for three reasons: to get better insight in the police domain; to be able to measure the first impression after an expert is told about an RM; and to get inspired what to focus on in the experiments.

Then, we performed a pilot experiment with nine students from the Radboud University. Finally, we performed the same experiment with five policemen, albeit with some adjustments. In the experiments, we collected data of participants interacting with an application simulating preventive frisking supported by a hard-coded RM and their experience of the RM. Where the experience is measured by the use of a survey. The fact that the RM is hard-coded means that the RM intervenes in the application with a pre-determined frequency and type of intervention depending on the specific case.

4.2 Expert interview

Because we are no experts in the police domain, it is necessary to communicate with and involve experts in the process in order to be able to make an RM that is experienced as useful and has an influence. An expert interview is a method of getting to know a particular field a bit better. It is a qualitative research method.

We were not that known in the police domain so we thought this would be a good way of getting into it. We performed this expert interview with a policeman in training from our network. Understandably, he did not want us to use his answers explicitly in our thesis. This is why we cannot include the authentic answer. However, the knowledge he gave us is used implicitly throughout this thesis and he agreed with us using a part of his answers explicitly, which can be found in chapter 5.

4.3 Experiment Design

4.3.1 Sub-questions

In constructing and doing the interview, we came to the conclusion that there are a number of questions to be asked about the setup of an RM. We can categorize these sorts of questions into five categories: goal (i.e. what and why), content (i.e. content of the reflective interventions), lay-out (i.e. how to present it to the experts), implementation (i.e. how is it implemented in real life?) and framework (i.e. how to achieve the goal?). The first category, 'goal', is treated in the chapters Introduction and Preliminaries. In this chapter we focus on trying to answer four sub-questions in the categories 'content' and 'lay-out'. The implementation and framework are not treated in this thesis and remains open to further research.

In the category content, we defined two interesting questions: "What type of reflective interventions do policemen prefer?" and "With what frequency should the RM pose a reflective intervention?". To label them, we specified them into two variables: type and frequency.

Next to these two variables, we also tried to see whether it is possible to answer one questions from the category lay-out: "Does the experience of interacting with a Decision Support System (DSS) influence the experience of the RM?" where experience of the RM is measured by the level of quality of the reflective interventions and the level of feeling criticized by the reflective questions and notifications of the participant. A DSS is a system which supports a decision-making process, hence the name. We did not actually include this in our application, though we told half the participants a story about this DSS selecting potential weapon carriers with the use of smart cameras. We included this question because we are curious whether users will experience the reflection machine in a different manner when being (falsely) informed about a decision support system.

4.3.2 Experiment flow

Before we dive into more details about the particular variables, we will explain the rough flow of the experiment. The framework of the application from the participants' point of view is as follows (also see Figure 4.1): first, the participants are presented a video clip of a few seconds. This video clip contains footage of several pedestrians on a square. Within this frame, a certain individual X is highlighted with use of a screenshot from the video. Here, the participants need to decide whether or not they want to frisk individual X, they can either answer 'yes', 'no', or 'I don't know'. There is also the opportunity to motivate the decision. After a button is pressed in order to continue, there are two things possible dependent on which case; either the participants continue to the next case or the RM intervenes. If an intervention takes place, this means that a reflective question or notification is presented. Dependent on which type and instance of the intervention, the participants may need to answer the question or press a button to continue. Then, in the last frame they receive the opportunity to change the initial decision and continues to the next case. See the appendix A.4 for screenshots of the application.

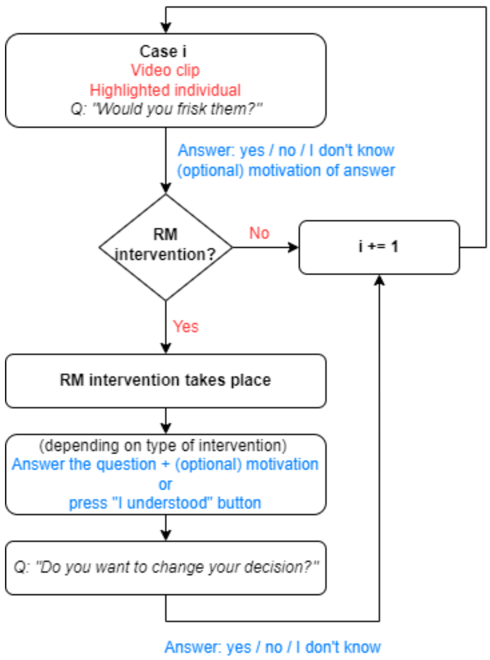


Figure 4.1: Flow chart from the user's point of view

This Figure displays the flow of the application from the user's viewpoint. It starts at the top with case i and continues downwards. Note that red indicates fixed aspects in a certain case i, while blue indicates the interaction of the user, e.g. giving an answer to the question.

4.3.3 Variables about content

For the variable type, we specified three types of questions that can be used to measure which one is experienced as most relevant: [1]:

1. Certainty based. This type of intervention aims to question the degree

[1]For all instances of the type of questions see appendix A.1

of certainty of the user's decision. It gives the users the opportunity to evaluate the certainty of their initial answer. There are three general instances of this type: *How certain are you about your decision (not) to frisk this person? Why?* where the question is answered with a rate bar on a scale from 1 to 5; *Are you sure about your decision (not) to frisk this person? Why (not)?*; and *Are you sure your focus will lead to an optimal decision? Why (not)?* where these questions are answered with a yes/no/I don't know.

2. Population based. This type of intervention aims to inform and make the user aware of the population in several aspects. For this type of intervention, it is required to have information from the population of the particular area. There are three general instances of this type: *Do you think this person deviates from the population in a particular aspect? Why (not)?* where this question is answered with a yes/no/I don't know; *Consider that [X] % of all weapons carriers of this population ...*; *Consider that [X] % of all people of the population walking past this spot ...* where these notifications are answered by a 'I understood' button.

3. History based. This type of intervention aims to inform and make the user aware of his past behavior regarding preventive frisking. For this type of intervention, it is important to note that it is required to keep track of particular attributes of the frisked individuals. We realized this by annotating the videos. Two general instances of this type are: *Consider that [X] % of all the people you decided (not) to frisk up to now, ...*; Up to now, you have decided (not) to frisk [X] % of all the people presented to you.

We came up with these types because this nicely increases in complexity of intervention. Certainty based is least informed and merely triggers the user to think about their degree of (un)certainty in their decision. Whereas population based as well as history based are informing the users about things they may have not taken into account in their decision and thus triggering the users to evaluate their reasons behind their decision. An advantage of categorizing these types of interventions is that it can be easily translated to another domain, like the medical field.

The frequency variable is divided into two instances: half of the cases are followed by an RM intervention; and all of the cases are followed by an RM intervention. This variable is measured within subject. The way this is exactly measured is explained in section 4.3.7.

4.3.4 Variable about lay-out

There is another variable from which we measured the external influence on the experience of the RM, we specify this as the variable 'DSS'. In both the student and the police experiment, half of the participants were told that we aimed to research the experience and effectiveness of a decision support system (DSS) while they are supported by our RM (see appendix A.2.2 and A.2.2). We gave them the false information that this DSS selects potential weapon carriers with the use of smart cameras and that the algorithms are not perfect which makes it essential to keep a human in charge: the participant. Conversely, the other half of the participants is uninformed about the DSS. They are told that we selected the individuals randomly ourselves (which is the case). Though, all participants are told about the support by our RM.

It is important to stress that this story about the DSS is not the truth. No system is used to select the individuals, we did this ourselves. However, we decided to implement this because in this way we might give the participant the feeling of criticizing an algorithm while feeling less criticized by the RM. Also, because they are busy thinking about the algorithm, they are more motivated to explain why the selected individuals do or do not make sense to them. And thus, we gain more insight in their reasons on why they decide whether or not to frisk an individual.

4.3.5 Participants

The students who participated in the pilot experiment are all from the Radboud University. Nine students participated. Whereas for the police experiment, we got five policemen from our network willing to participate in this study. Three of these policemen actually had experience in the field of preventive frisking, whereas the other two did not.

The participants were told that there is a high crime rate in population Z. Where 70 out of 100 people own a weapon, which motivated the government to put the weapon rate down by commencing a project in preventive frisking. This crime rate is made up to stimulate the participants to frisk people. All participants were informed about the study upfront (see A.2.2) and debriefed (see A.2.3) afterwards. Each participant filled out a consent form (see A.2) before participating in this study. We ensured that all experiments are performed on a laptop in a silent and non-distracting environment, either on campus or on the police station.

4.3.6 Video clips

For the video clips, we used the data from WILDTRACK. This is an open-source dataset containing 35 minutes of footage of seven different cameras placed in a different viewing angle on the square. This new multi-camera dataset was acquired using seven high-tech statically positioned cameras with overlapping fields of view on the same square. The data acquisition took place in front of the main building of ETH Zurich, Switzerland, during nice weather conditions.

We selected fifty individuals from three different cameras. We ensured that these individuals are clearly visible and that fifty percent of the individuals are men, whereas the other fifty are women. We also made sure to select people walking alone as well as in a group. Then, we annotated the videos by describing the following attributes of every individual: pace; nr. of people person is with; facial expression; focus; estimated age; clothing; sex; accessories; other outstanding aspects. These can be found in the Github repository. Despite the fact that we annotated all fifty videos, we only used a fixed set of twenty videos for the eventual experiment. More details on what videos are used can be found later in this chapter.

4.3.7 Student design

Our experiment consists of four different designs (see Figure 4.2). Two of them (a1 and b1) include informing the participant about a decision support system, whereas the other two (a2 and b2) do not. The second distinguishing is a split in designs to overcome the interference of the possibility of people being tired at the end of the experiment. This makes our experiment design consists of four distinct variants:

- a1 Following order of parts: part A - part B and informed about DSS
- a2 Following order of parts: part A - part B and not informed about DSS
- b1 Following order of parts: part B - part A and informed about DSS
- b2 Following order of parts: part B - part A and not informed about DSS

The experiment for the students consists of six blocks divided over three parts. Every block consists of four cases. That makes a total of twenty-four cases. The cases have a mix of men and women in a 50/50 ratio per block, this means that in every block two women and two men were shown. All cases are fixed, i.e. for each participant it holds that for each case they are presented the same video clip as the other participants. Though, while the categories of the RM interventions are fixed, the instances are randomized (i.e. it is possible that for some case i one participant is shown question 1 from the certainty based interventions, while another is shown question 3 of this same category).

In block 1, the variable frequency instance is 'half of the cases' and in block 2 'all of the cases'. The result of this is that in block 1, the participants got to see four cases in which two of them are not followed by an RM intervention, whereas the other two do. In blocks 1 and 2, the variable type is fixed after being randomized i.e. for each case in these blocks we have randomly assigned one of the type instances, while ensuring that for every participant this is the same. Thus, for example in case 2 of block 1, the video clip is always followed by an RM intervention where the type is always population based with the instance: 'Consider that 90% of all people of the population walking past this spot are carrying luggage', despite of the fact that this instance is assigned randomly.

For each block 3, 4 and 5 the participants are presented with solely one type in all four cases; types certainty based, population based and history based respectively. In all these blocks, the variable frequency is set with instance 'all of the cases' which entails that every case is followed by an intervention such that there are as many RM interventions as possible to be able to measure the experience of the variable type.

The last block consists of four cases where not only the cases and the RM interventions, but also every answer to each question is fixed. In the survey we would ask the participants whether they think the person behind the answers (which was actually no one as we came up with the answers ourselves) experienced the reflective interventions. We included this block to measure a potential effect of the bias blind spot, however later we realized this is not the way so we omitted this block and did not take this into account in the results.

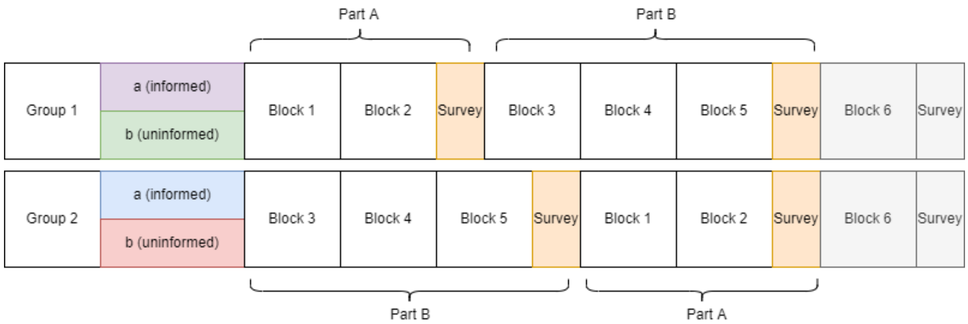


Figure 4.2: Experimental setup

This Figure emphasizes the setup which is split in four groups. Group a1 and a2 are informed about the DSS, whereas Group b1 and b2 are not. Group 1 received the parts in the order A $\longrightarrow$ B, whereas Group 2 in the order B $\longrightarrow$ A. The exact same colours are used as in the excel sheets to be consistent, these sheets can be found in the GitHub repository (`https://github.com/Hann-v0/ReflectionMachine.git`). We also indicated that Block 6 and survey 3 are not included in the analysis by colouring them grey.

4.3.8 Policemen design

A similar division is made for the policemen design as for the students. However, for this design we omitted the last block and survey, as explained above. Consequently, we only had five blocks which makes up a total of twenty cases. Another change is that this time, we also set the instances of the RM intervention (i.e. if one participant saw the first question from the category certainty based for some case, all of the other participants received this very same intervention).

4.3.9 Evaluating the experience of the reflection machine

The evaluation is done with the use of a survey. The survey setup is in the form of a 5-point Likert scale with the specific label values of the rating depending on the question. After each part (i.e. after block 2, 5 and 6, see Figure 4.2) a survey is filled in by the participants in which they were presented a total of eight or nine statements which they had to give a rating. The number of statements depends on the completed part of the application and whether they had the DSS condition or not. For example, statement 9 saying: 'You saw the following reflective questions / notifications. Please rank them in order of relevance.', is asked solely after the participant completed block 5. After each statement, there was a possibility to motivate the answer. The statements can be found in appendix A.3.

Statement 9, 4, 6&7 and 8 are used to answer sub-questions 1 to 4 respectively. Statement 2 is useful to be able to analyze the effectiveness of the RM, but this is outside the scope of this thesis. Some statements (1,3,5) are put in only to distract the participant from the actual goal of the experiment to stimulate the participants to give respond elaborately in the application. For example statement 5 saying: 'I think the concept of the algorithm could be helpful in the field of preventive frisking' is included (only in the dss condition) to draw attention to the DSS.

4.4 Expectations

4.4.1 Content variables

Type of RM interventions

We expect that the participants would consider the population interventions as most relevant and the certainty variable as most irrelevant. That leaves the history interventions in between. We believe this will be the case because certainty based questions do not really encourage considering the problem at hand from a different viewpoint, it merely makes the user reconsider. Whereas population based interventions actually inform the users about the population and potentially encourages them to take other aspects into account. We argue that history based interventions are more easily experienced as annoying than helpful.

Frequency of RM interventions

In general, we expect that participants would like less interventions because of two reasons. First, because it may get annoying in a while and second because it may be experienced as being continually critiqued. We included two statements in the survey to question quality and criticism.

4.4.2 Lay-out variable

Our expectation is that the DSS influences the experience of the RM in a positive manner. We think that by giving the participants the task to focus on the performance of the algorithm suggesting individuals from the video clips, the focus shifts from being questioned (by the reflection machine) to questioning (the dss). Which gives the participant the feeling of being an expert rather than a trainee. We argue that in this way, the total experiment is associated with a more positive feeling resulting in a more positive association regarding the reflection machine.

4.5 Evaluation of study design

We came up with this analysis design in order to do a combination of qualitative and quantitative analysis. It was out of the scope of this thesis to gather more policemen willing to participate in this experiment in order to actually do a statistically significant quantitative research.

Another opportunity would have been multiple interviews or a focus group to make it more of a qualitative research. However, a disadvantage of qualitative research is the large room for interpretation of the researcher. Next to that, in the experiment, the participants interact with a computer which makes it easier to be honest about your answers due to the absence of the social factor. Furthermore, the experimental setup gives us the ability to visualize the results in a neat way, which is also harder with qualitative research. Fortunately, because of possibility for the participants to motivate their decisions, we can dive deeper into the reasons behind the answers in an exploratory and more qualitative manner.

Chapter 5

Results

5.1 Expert interview

We want to highlight some aspects from the expert interview with approval of the expert. We want to emphasize that he is a policeman in training and is not a representative for all Dutch policemen. Furthermore, we have to note that at this point we still used the term falsification machine instead of reflection machine. In order to prevent confusion, we will use the term FM in this section.

The domain expert had his doubts about the effectiveness of a system like the FM. He argues that it has a high chance of being a superfluous system because in police training, they are already educated to reflect on themselves and their colleagues with questions like "Why did you act the way you did?" and "What could you have done better?". There is a large emphasis on reflecting during the training partly to minimize the inclusion of personal bias. This is supported by the webpage of firm requirements to be allowed to work at the police of the Netherlands [17], they state: "And as a police officer or woman you act on the basis of the norms and values of the police organization, you must set aside your own opinion and prejudices. Are you able to treat everyone equally without discrimination?"[1]. Obviously, when they complete the police training it is expected they are able to reflect upon themselves and coworkers and professionally handle situations. Our expert has doubts about whether an FM is able to make a change in the few cases where reflection is not included in acting as much as desired and taught in training. The first impression of the expert when we told him about the FM was that it felt like a criticizing machine. He argues that it may happen that users get annoyed by the questions. We argue that this may be the case because of using the word 'falsification', which has a more critiquing

[1]Translated from Dutch: "En als politieman of -vrouw handel je vanuit de normen en waarden van de politieorganisatie, je dient je eigen mening en vooroordelen opzij te zetten. Ben jij in staat iedereen gelijk te behandelen zonder onderscheid te maken?"

insinuation, instead of 'reflection' with a more supporting connotation.

After asking the expert whether he recognized the problem of distrust in police, he mentioned the opposite: he does experience a large amount of trust from the general public. He argues that in the small group experiencing distrust in police, it would mainly be caused by misinformation, which an FM cannot take away. In other words, the small lack of trust of citizens in police may not be solved by supporting the police with an FM.

However, if we want to continue working with an FM, he states that the most important aspect of the system should be to make sure to not be experienced as criticizing, but as supporting.

5.2 Experiment

Before we dive into the results of the experiment, we want to stress that we made an important assumption over the data. We cannot be certain that the difference between "very bad" and "bad" is the exact same difference as "good" and "very good". Though, in this data analysis, we will explicitly make the assumption that each step in between two answers on a rating scale represents the exact same difference.

5.2.1 Experienced relevance of type of RM intervention

The first sub-question we will answer is: "What type of reflective interventions do policemen experience as most relevant?". In order to do this, we use the data of the participants responding on the statement: "You saw the following reflective questions / notifications. Please rank them in order of relevance." This question has only appeared once for each participant; in the survey after part B, where we made sure that the participant had seen all possible interventions. In this question, all reflective interventions are shown to the participant (see appendix A.1) labelled by C1, C2, C3, P1, P2, H1, H2. Where C, P, H mean Certainty based, Population based an History based respectively. The 1, 2, and 3 are the indicators of a particular instance of that type of intervention (e.g. C1 refers to "Are you sure you want to frisk this individual? Why (not)?"). To be clear, the labels are not presented to the participant, solely the interventions. The participants scored the relevance of all instances where 7 is the highest score and 0 the lowest. We first analyze the student data and then we do the same for the police.

Relevance ranked by students

Although we only have nine participants, we will average the scores for each intervention instance over the participant such that we can work with a 9 x 1 (instances x averages) matrix instead of a 9 x 9 (instances x participants) matrix. For the original data, we refer you to our GitHub repository (`https://github.com/Hann-v0/ReflectionMachine.git`). In Figure 5.1, you can see clearly that, on average for the students, C1 is experienced as most relevant, while P1 is experienced as least relevant. We also included the standard deviation in order to see whether the participants generally agree with each other. For most of the instances, we see a large standard deviation indicating that the experienced relevance may largely differ per participant. Now, when we cluster all intervention instances to form the three categories, we construct Figure 5.2. Here we can see that on average, students experienced type C as most relevant.

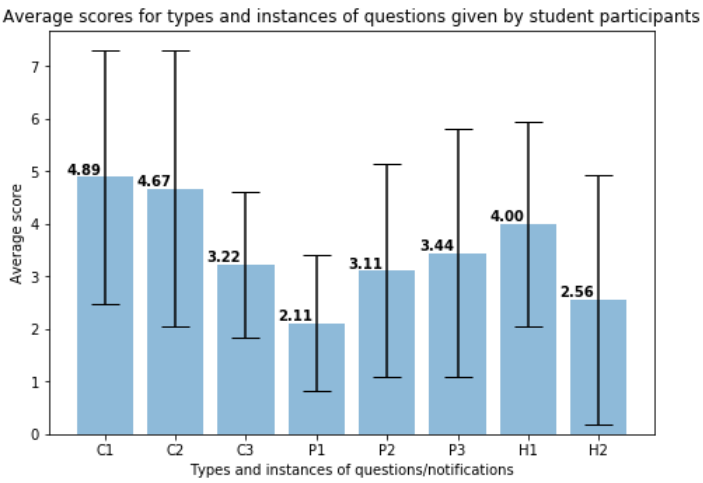


Figure 5.1: Relevance of distinct interventions ranked by students

The relevance is measured by scoring the most relevant intervention with 7 points, up to 0. For each intervention, the scores are added and averaged out over the number of students. The numbers indicates the averaged score. We also include the standard deviation to see the spread in scoring the intervention. On average, C1 is given the highest score and P1 the lowest.

Relevance ranked by police

For the police, we performed an identical analysis; first, we took the average for every instance, and then we clustered the instances of the types together. In Figure 5.4 you can see the result. Here again, we see that on average C1 has gotten the highest score (i.e. 5.2) and is thus experienced as most relevant, while P1 is experienced as least relevant with an average of 1.2 points. In Figure 5.4 we see a similar trend as for the students. We can see that again type C is experienced most relevant and type P is least relevant. However, the difference between C and H is not as large as for the students.

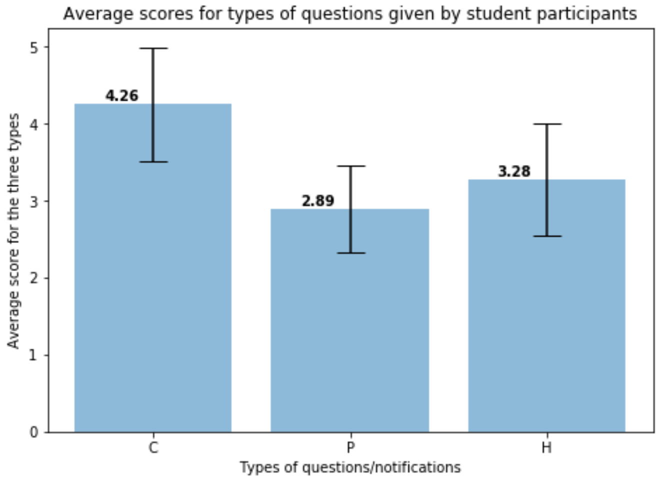


Figure 5.2: Relevance of distinct interventions ranked by students - clustered

Category C (certainty based interventions) accounts for the average of C1 to C3; P (population based interventions) for the average of P1 to P3; H (History based interventions) for the average of H1 to H2. On average, category C is experienced as most relevant, while P is experienced as least relevant.

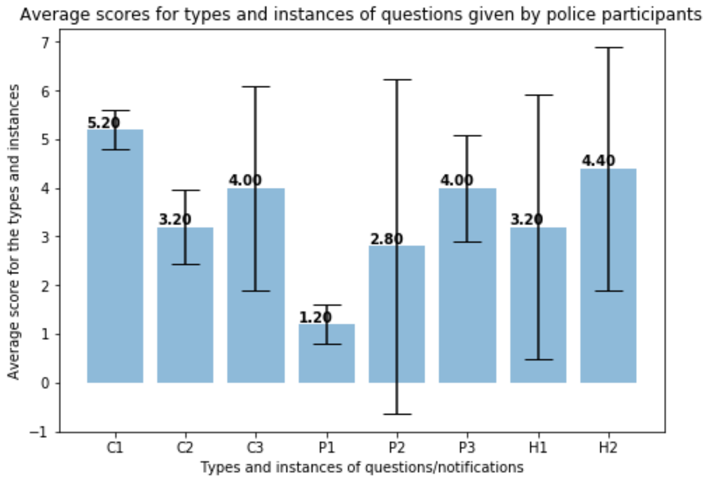


Figure 5.3: Relevance of distinct interventions ranked by policemen

The same scoring technique as for the students is used for the data of the policemen. For each intervention, the scores are added and averaged out over the number of policemen. The numbers indicates the averaged score. Again, on average, C1 is given the highest score and P1 the lowest.

5.2.2 Frequency of RM interventions

The second sub-question is: "How often should the RM pose a reflective intervention?". To answer this question, we focus on the survey question

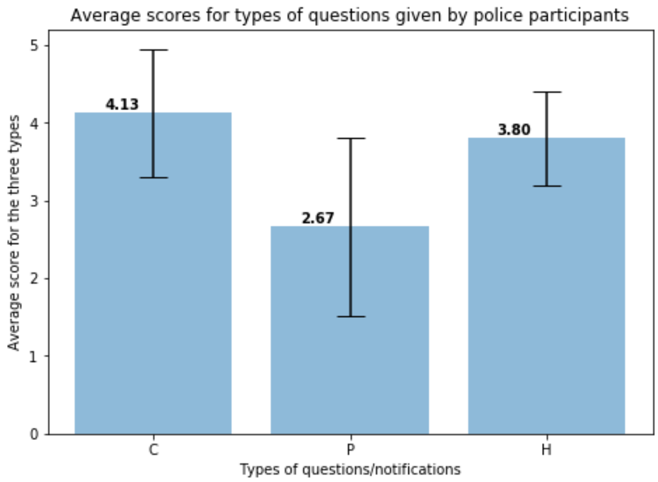


Figure 5.4: Relevance of types ranked by policemen - clustered

Category C (certainty based interventions) accounts for the average of C1 to C3; P (population based interventions) for the average of P1 to P3; H (History based interventions) for the average of H1 to H2. On average, category C is experienced as most relevant, while P is experienced as least relevant.

"How do you experience the amount of reflective questions and notifications?". The participants had to rank this question with the two border values being: "the less the better" and "the more the better". This question is asked twice to each participant, once after each part (part A and B, see Figure 4.2). So we have a data matrix of 9 x 2 for the students and 5 x 2 for the police. Again, we will first analyze the answers of the students.

Frequency preferred by students

We listed all answers in Figure 5.5. Within subjects there are some changes: a few participant switched from "I like more" to "Neutral" and vice versa over the different parts (i.e. sp2, sp5, sp7 and sp8). Some participants indicated that they did like particular interventions, while they did not like others. One of these participants (sp4) argues that: "I think the amount of reflective questions and notifications are just enough. They make you think about your decision and are not annoying right now. I think it is good that after every decision you are prompted with a question/notification so that it does not become an automatic process of just saying yes I would frisk this person or no I would not frisk this person. If there would be more questions/notifications I think it would become cumbersome as you already have thought about your decision twice.". Some even motivate their rating by saying that they would like more interventions per case, while others (e.g. sp3) indicate that they would not miss them if they were not there.

```
>>>
Survey Question: ' How do you experience the amount of reflective questions and notifications?

Participants    Survey part A          Survey part B
-------------   -------------------    -------------------
sp0 (B->A)      I like more;           I like more;
sp1 (B->A)      The more the better;   The more the better;
sp2 (A->B)      Neutral;               I like more;
sp3 (B->A)      I like more;           I like more;
sp4 (B->A)      Neutral;               Neutral;
sp5 (A->B)      I like more;           Neutral;
sp6 (A->B)      Neutral;               Neutral;
sp7 (A->B)      Neutral;               I like more;
sp8 (A->B)      I like more;           Neutral;
```

Figure 5.5: Responds from student on statement 4 of the survey

The responds to a question from the survey regarding the amount of reflective interventions. "sp" indicates that it was a student participant, the number indicates the particular participant. After this, we indicate in which order they took the different parts. In column 2 and 3, you can see the answer given. No one ranked this below Neutral (i.e. "the less the better" or "I like less")

Frequency preferred by police

The responses of the policemen can be seen in Figure 5.6. Within subjects there is only one change: participant 0 switched from "The less the better" to "Neutral". He also motivated his first answer in both parts, in part B: "My basic training taught me not to deviate too much from your first thought. We often have to make quick decisions in our work.", while he motivated his second answer in part A with "Sometimes it seemed like other questions that I had to think about, plus I was a little less sure if I needed to check on someone.". Another participant (pp4) motivated his "The more the better" answer with "In this way you get a more specific answer."

In between subjects, there is not much variation. Participants either chose the more the better or neutral. Only one of them (pp0) chose the less the better, as explained above.

5.2.3 Influence of being informed about a DSS

Does the belief of interacting with a DSS influence the experience of the RM? Here, the experience of the RM is measured by the level of quality of the reflective interventions and the level of feeling criticized by them. We measured this variable by making use of the responses to two survey statements: "How do you experience the quality of the reflective questions and notifications?" and "I felt criticized by the reflective questions and notifications". Again, we did this for both police and students.

It is important to stress again that the instances of the reflective interventions were randomized for the students, whereas they were set for the

```
>>>
Survey Question: ' How do you experience the amount of reflective questions and notifications?

Participants     Survey part A          Survey part B
--------------   --------------------   --------------------
pp0 (B->A)       Neutral;               The less the better;
pp1 (B->A)       Neutral;               Neutral;
pp2 (B->A)       The more the better;   The more the better;
pp3 (A->B)       The more the better;   The more the better;
pp4 (A->B)       The more the better;   The more the better;
```

Figure 5.6: Responses from students on statement 4 of the survey

The responds to a question from the survey regarding the amount of reflective interventions. "pp" indicates that it was a policeman participant, the number indicates the particular participant. After this, we indicate in which order they took the different parts. In column 2 and 3, you can see the answer given. Only one of them ranked below Neutral (i.e. pp0-surveyB with "the less the better")

policemen. This likely results in a very different experience. Nonetheless, we will show the results of both the students and the policemen.

How do you experience the quality of the reflective questions and notifications?

This question is rated on a 5-point Likertscale with 'very bad' and 'very good' as most extreme values. In Figure 5.7, a table is shown where all answers to this question of each participant are shown. Next, we split the data of the participants in two conditions: (falsely) informed about the DSS and (truly) uninformed about the DSS. From now on, we will label these the informed and uninformed conditions respectively. The results can be found in Figure 5.8, where red indicates informed condition, whereas blue indicates the uninformed condition. It may be due to the small number of participants, but we do not see a large difference between them.

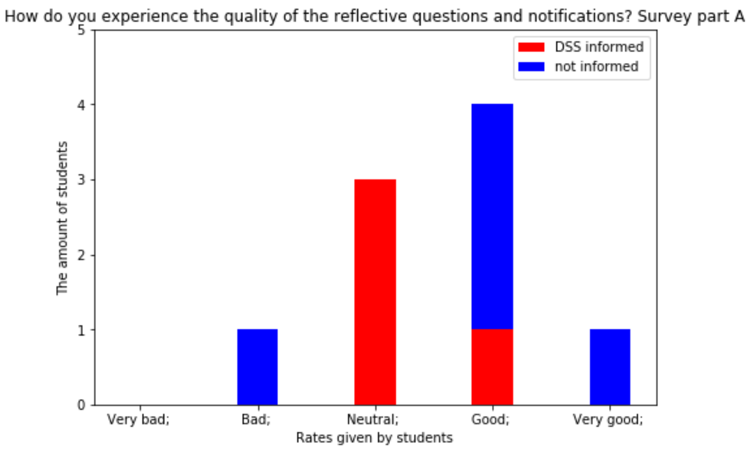
In Figure 5.9, a table is shown including the police participants. We analyzed the same statement for the policemen as for the students and got the results shown in Figure 5.10. Three of them motivated their answers. One of them (pp2) argues that the interventions are sometimes containing a suggestive tone. Participant 0 reasons that: "It does not interest me that much. I think a person is suspicious or not, then I don't think about percentages or anything like that." and a participant (pp1) indicates that he is in doubt about his rating.

```
>>>
Survey Question: ' How do you experience the quality of the reflective questions and notifications?

Participants   Survey part A   Survey part B
-------------  --------------  --------------
sp0 (B->A)     Good;           Good;
sp1 (B->A)     Very good;      Very good;
sp2 (A->B)     Good;           Bad;
sp3 (B->A)     Neutral;        Good;
sp4 (B->A)     Good;           Good;
sp5 (A->B)     Good;           Neutral;
sp6 (A->B)     Neutral;        Neutral;
sp7 (A->B)     Neutral;        Bad;
sp8 (A->B)     Bad;            Bad;
```

Figure 5.7: Responses from students to survey statement 6 in a table format

This table represents the answers to statement 6. "sp" means that it was a student, the number indicates the particular participant; A and B specify the order. The columns represent the given answers. It ranges from 'Bad' to 'Very good'.



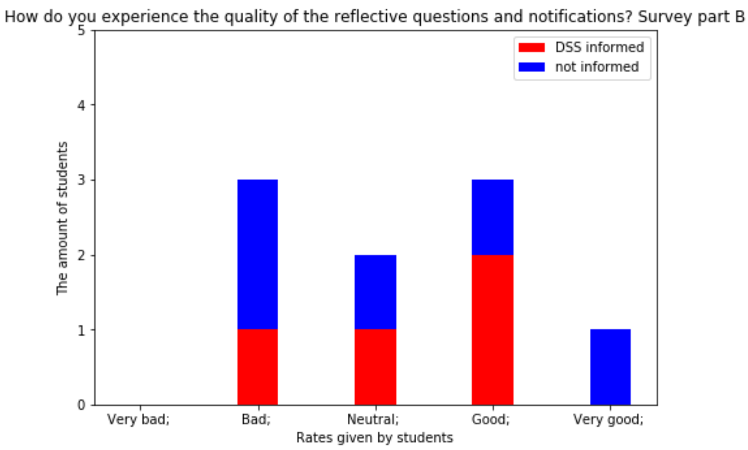


Figure 5.8: Responses from students to survey statement 6

Here, we see the distinction of the two conditions (informed and uninformed) presented in a bar plot. We do not see a clear difference.

```
>>>
Survey Question: ' How do you experience the quality of the reflective questions and notifications?

Participants    Survey part A    Survey part B
--------------  ---------------  ---------------
pp0 (B->A)      Good;            Bad;
pp1 (B->A)      Good;            Good;
pp2 (B->A)      Good;            Very good;
pp3 (A->B)      Good;            Good;
pp4 (A->B)      Neutral;         Neutral;
```

Figure 5.9: Responses from policemen to survey statement 6 in a table format

This table represents the answers to statement 6. "pp" means that it was a policeman, the number indicates the particular participant; A and B specify the order. The columns represent the given answers ranging from 'Bad' to 'Very good'. It is mainly answered with 'Good'.

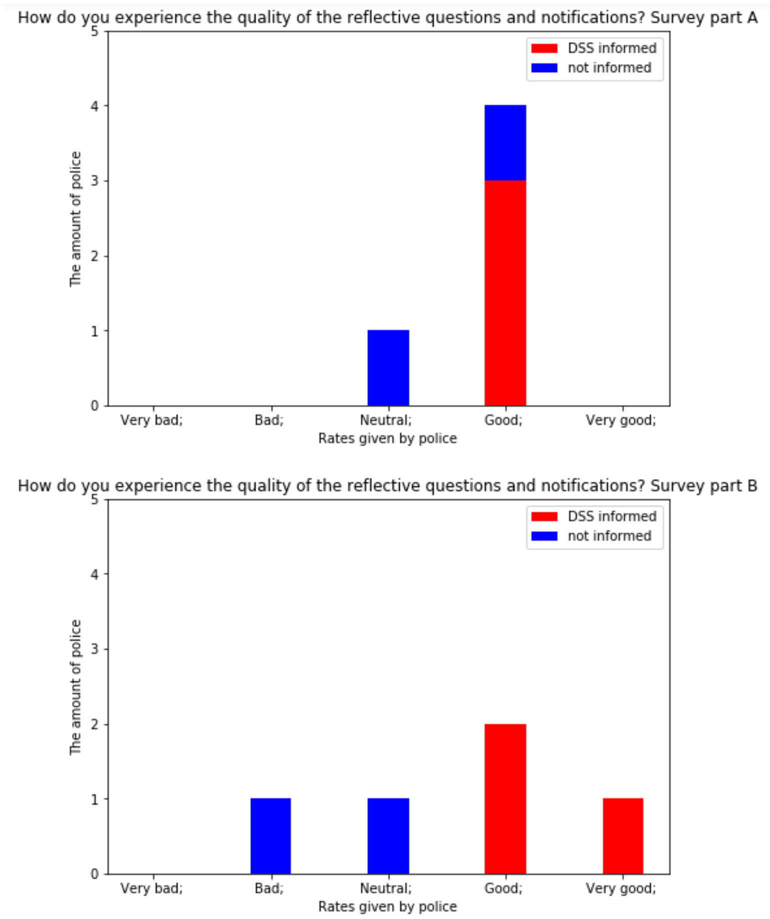


Figure 5.10: Responses from policemen to survey statement 6

Here, we see the distinction of the two conditions (informed and uninformed) presented in a bar plot. We do identify a potential difference between the two conditions, especially in part B.

I felt criticized by the reflective questions and notifications

This statement is rated on a 5-point Likertscale ranging from 'Very bad' to 'Very good'. In Figure 5.11, a table is shown where all answers to this question of each participant are included. In Figure 5.13 the results are shown where we split it into the two conditions. The spread is large in the informed, it really differs per participant. Only participant 0 motivated the given answers 'Agree' to 'Neutral'. In the survey after part B the motivation was: 'A little bit, which I also thought was funny. It shook me awake for a while that I do have a certain bias myself.' , while after part A, it was: 'This time less than last time. But it does keep you sharp.'. Now, we look at the uninformed condition. There is no difference between in bar plots in the two parts. All participants stand by their first indication and they generally do not feel criticized. One participant (sp4) motivated both his answers with 'The notifications I received were very much in line with what I thought I had answered so I was not surprised by them.', whereas another participant (sp5) does feel criticized, indicating: 'I especially felt criticized by the notifications, because they almost seem to assume that I am biased for, for example, ethnicity or age. I probably am too, but that's not very nice to hear.'. We even got some insight in the different types of interventions (sp2): 'I don't care about ratio's of my own work, only statistical reminders of who is most likely to have a weapon'.

In Figure 5.13, it is made clear that there is literally no difference between the two parts. We also see no noticeable difference between the informed or the uninformed condition.

The table including the police participants is shown in Figure 5.12. For the police, we see no indication of any relation between the informed and uninformed condition (see Figure 5.14). However, due to the number of participants and the similarity of the data, this is no firm conclusion. To be able to really say anything meaningful about this, we suggest further research.

```
>>>
Survey Question: ' I felt criticized by the reflective questions and notifications

Participants    Survey part A       Survey part B
-------------   -----------------   -----------------
sp0 (B->A)      Neutral;            Agree;
sp1 (B->A)      Strongly disagree;  Strongly disagree;
sp2 (A->B)      Strongly agree;     Strongly agree;
sp3 (B->A)      Strongly disagree;  Strongly disagree;
sp4 (B->A)      Disagree;           Disagree;
sp5 (A->B)      Neutral;            Neutral;
sp6 (A->B)      Neutral;            Neutral;
sp7 (A->B)      Agree;              Neutral;
sp8 (A->B)      Strongly disagree;  Strongly disagree;
```

Figure 5.11: Responses from students to survey statement 7 in a table format

This table represents the answers to statement 7. "sp" means that it was a student, the number indicates the particular participant; A and B specify the order. The columns represent the given answers ranging from 'Strongly disagree' to 'Strongly agree'.

```
>>>
Survey Question: ' I felt criticized by the reflective questions and notifications

Participants    Survey part A       Survey part B
-------------   -----------------   -----------------
pp0 (B->A)      Disagree;           Disagree;
pp1 (B->A)      Disagree;           Strongly disagree;
pp2 (B->A)      Strongly disagree;  Strongly disagree;
pp3 (A->B)      Strongly disagree;  Strongly disagree;
pp4 (A->B)      Strongly disagree;  Strongly disagree;
```

Figure 5.12: Responses from policemen to survey statement 7 in a table format

This table represents the answers to statement 7. "pp" means that it was a policeman, the number indicates the particular participant; A and B specify the order. The columns represent the given answers. It ranges from 'Strongly disagree' to 'disagree'.

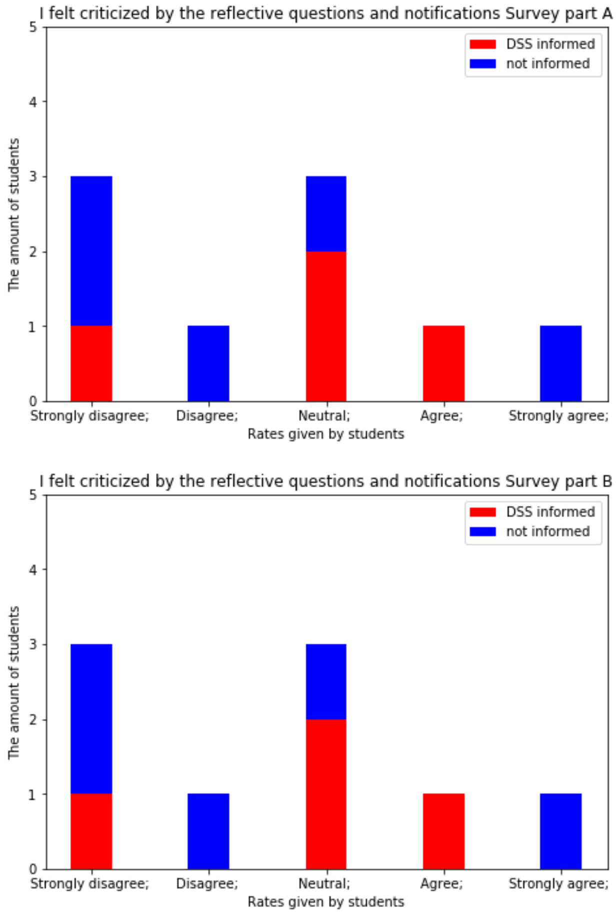


Figure 5.13: Responses from students to survey statement 7 split over DSS condition

Here, we see the distinction of the two conditions (informed and uninformed) presented in a bar plot. Whereas we do not see a clear difference between the two conditions, it is noticeable that there is no difference in between parts.

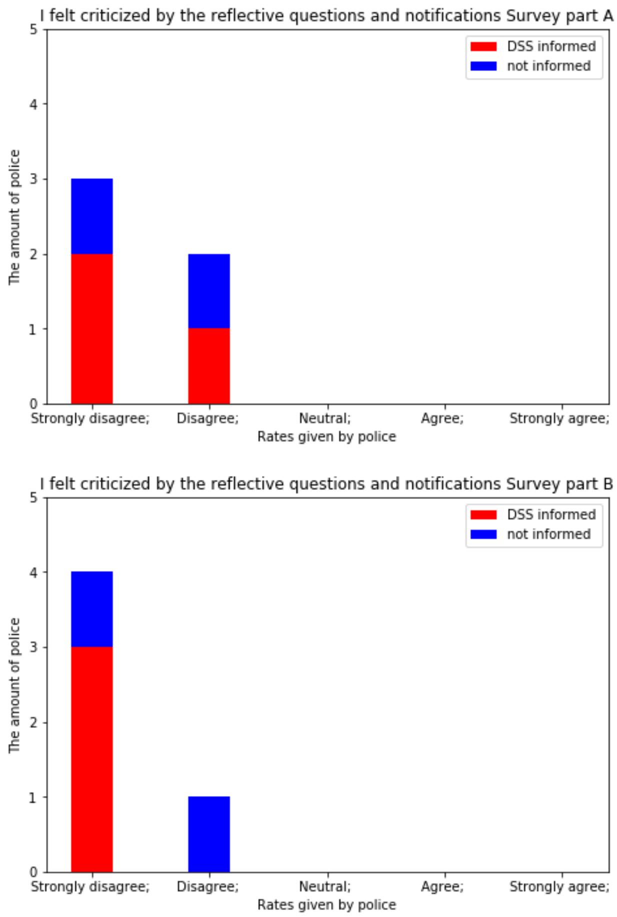


Figure 5.14: Responses from policemen to survey statement 7

Here, we see the distinction of the two conditions (informed and uninformed) presented in a bar plot. We do not see a clear difference between the two conditions, nor parts.

5.2.4 Exploratory Analysis

In this paragraph we will explore the data obtained from the application. There are six variables for each case for all participants (see Figure 4.1): whether they want to frisk or not; (optional to user) motivation to why they (do not) want to frisk; (in some cases) an RM intervention; (in some cases) response to RM intervention; (optional to user) motivation over response to RM; (in cases where RM intervenes) whether or not they want to change decision. While this does not really have anything to do with the original aim of this thesis, we added this paragraph because this data contains lots of interesting information from the experiment. It gives us valuable information in the reasons behind the frisking and it gives us insight in whether participants want to change decision after interacting with the RM. We mainly focus on the application data of the policemen as we mostly saw comments in the data of the students that we think are not very useful (e.g. a motivation to why frisk someone: "A smoker generally cares less about peoples health and well-being").

First of all, it is interesting to note that there are four out of twenty cases in which each policeman made the same decision in whether to frisk or not: case 5, 8, 11 and 17 (see Figure 5.15). As we annotated the video clips upfront and we gathered the motivations, it is easy to see why: one of the highlighted individuals (case 5) changes direction after seeing the camera and all of the participants indicated this as the reason why they wanted to frisk this individual. Another individual (case 11) wears a hoodie and sunglasses and walks relatively fast. It is interesting to note that not only all policemen, but also every student wanted to frisk this person; this is the only individual for which this is the case. The other two cases in which every policeman made the same decision (case 8 and 17) are where they do not want to frisk this person. The reasons behind it generally include something like "no deviant behavior". In Figure 5.16 we listed the amount of frisks per participant, which largely differs. It is funny to see that for participant 3 the only frisks he performed are the ones every other participant chose to frisk as well.

Another interesting fact is that in only four cases there was a participant who changed his decision after the RM intervened. However, we think one of these cases is a misunderstanding: in case 11, participant 1 changed his decision from yes to no after stating in the RM intervention asking him about his degree of certainty that he was 4/5 certain about his decision. This is no coherent pattern. In case 10, participant 3 changes his decision from 'I don't know' to 'Yes' because he would "call him out" on his behavior (i.e. this individual shows his middle finger towards the camera). Some other case shows us the uncertainty of participant 3 who initially was not sure whether to frisk or not, motivated by "he looks back, and looks like he's yelling at someone else. Who is that and what is his attitude/behaviour?",

	p0	p1	p2	p3	p4
0	Yes	No	Idk	No	No
1	No	No	Yes	No	No
2	Yes	No	No	No	No
3	Yes	No	No	No	Yes
4	No	No	Yes	No	Yes
5	Yes	Yes	Yes	Yes	Yes
6	No	No	Yes	No	No
7	Yes	Yes	No	No	Yes
8	No	No	No	No	No
9	No	No	Yes	No	No
10	Yes	No	No	Idk	Yes
11	Yes	Yes	Yes	Yes	Yes
12	No	No	Yes	No	Yes
13	Yes	No	Yes	No	No
14	No	No	Yes	No	No
15	Yes	Yes	Yes	Idk	Yes
16	No	No	No	No	Yes
17	No	No	No	No	No
18	No	No	Yes	No	No
19	No	No	No	No	Yes

Figure 5.15: Policemen wanting to frisk

The columns represent the police participants and the rows represent the cases. The data indicates whether the participant wanted to frisk a particular individual or not.

	Yes	No	Idk
p0	9	11	0
p1	4	16	0
p2	11	8	1
p3	2	16	2
p4	10	10	0

Figure 5.16: Police participants rates

The columns represent the answer options to the question whether the participants wanted to frisk or not. The rows represent the participants. We notice that participant 3 did only frisk 2 participants, when we look at Figure 5.15 we see that these cases are the cases in which everyone frisks this individual.

and eventually was not sure on whether to change his decision. Case 1 shows an individual walking down the stairs without anything noticeable about this video clip, but after indicating that "Consider that 95% of all weapon carriers of this population are carrying luggage." the initial decision of participant 3 not to frisk is changed to yes.

The last noteworthy aspect of the data is that the motivations behind the reason of frisking are suitable for clustering. After inspecting the motivations of the possible answers participants gave, we divided it into 6 distinct clusters why people chose to frisk an individual; and 5 distinct clusters why people chose not to frisk an individual. In some cases, participants gave multiple reasons why they wanted to (not) frisk a particular individual. The 6 clusters with reasons to frisk are: 1) distinct pace; 2) individual is alone; 3) suspicious focus; 4) distinct use of -or attention towards- accessories; 5) deviant behavior; 6) other, e.g. information told upfront or during experiment, gut feeling, it will not hurt to do it, etc. The 5 clusters with reasons not to frisk are: 1) not the type to be carrying a weapon; 2) factual description of the situation (i.e. without any conclusion); 3) no deviant behaviour; 4) cannot find a specific reason to frisk; 5) individual seems to be goal-oriented. We performed no further analysis on this data because of time constraints. However, we paved the way for potential future research.

Chapter 6

Conclusions

This research aimed to examine the functional preference of policemen concerning an RM in a preventive frisking context. Based on qualitative analysis from an expert interview we can conclude that the first impression of a former called falsification machine is potentially negative. We also experienced that in conversing with the policemen of the experiment upfront. However, this does not imply that an a reflection machine should not be implemented in the police domain; it is merely a first impression. By analyzing the experience and functional preference of policemen after interacting with an application simulating preventive frisking supported by a reflection machine, we can conclude that our participants generally like the reflective interventions in this setup.

We particularly focused on the experienced relevance of different types of reflective interventions. We divided the types of interventions into three categories with different levels of information included in these interventions. When asking the participants to rank them in order of relevance, it becomes clear that on average, for both students and policemen, they experience the certainty based interventions as most relevant. We conclude this may be the case because of a possible increase in confidence in their actions and reasons behind their actions. This could be a sign of the confirmation bias, where people tend to hang on to favoured hypotheses and become more confident in their own opinion because of it. This is a conflict between two of the goals of an RM, namely on the one hand decreasing bias awareness while on the other hand increasing confidence. However, the main goal is increasing bias awareness so this is contrary to the total RM goal. We suggest further research in the effectiveness of a potential reflection machine, where the different effectiveness of the different type of interventions are considered. We expect this is also helpful in order to be able to further specify the design choices that have to be made regarding the functional preference.

The results regarding the experience of the number of reflective interventions, i.e. participant indicate that in general that more interventions are desired, is in general a positive boost for the RM. Although, one policeman indicates that in training he is taught to listen to a first thought rather than deviating too much from it, we still conclude that in general, the police liked the reflective interventions. Please note however, that this is in a very different setting from real life and they were only asked to interact with it once. This is no guarantee that they would like to interact with it in a more regular way. However, as de Graaf states in her research, it may not be desired to implement it in a real life setting but rather as a training tool [7].

Whether the user of this application is informed or uninformed about a non-existing DSS may have a positive effect on the experienced quality of the reflective interventions, while we found no indication for a correlation between being informed about the DSS condition and the level of feeling criticized for both the students and the policemen.

In short, we can conclude that the hard-coded reflection machine in our application is experienced in a positive way. On top of that, although more research is definitely required in this field, our data supports the idea that an RM is able to be experienced as a supporting tool in a preventive frisking context and may achieve the desired goals, namely increase bias awareness, transparency, meaningful human control and confidence. We believe this is possible by ensuring that the level of experienced criticism is as low as possible, while the level of experienced quality is as high as possible.

Chapter 7

Discussion

The results indicate that in this setup participants generally like the interventions and mainly the certainty based interventions. We see several motivations in the survey supporting this. However, there are some limitations to this research. First of all, for both the student and the police experiment, the number of participants is not high enough to draw firm conclusions. This does not mean this entire research is redundant because it does contain data that supports the idea that the concept of an RM could have benefits in the police domain and we identified a possible functional preference to our defined certainty based interventions. In order to examine what the number of participants should have been to draw a firmer conclusion, we wanted to perform a power analysis where we would determine the effect size by making use of the pilot experiment. We concluded that in this particular research it is not appropriate to use power analysis where you can set two variables (i.e. significance and statistical power), determine a third (i.e. effect size) and calculate the fourth (i.e. sample size) using software tools such as G* power. This is because we use a more qualitative within-subject design where we explore experience rather than measure effect. This makes it out of place to scrutinize an effect size and thus do a power analysis. Regarding a next step in research of an RM in a preventive frisking context we suggest to measure the effectiveness of an RM, both short and long term. Here, we advice to assign one control group performing preventive frisking and one experimental group doing this while interacting with an RM. In such a framework, it is easily determined what the power of this research is.

Secondly, we explored the experienced relevance of different type of interventions. In examining this, we are not sure whether we really measured the relevance or the preference. For example, questioning the level of certainty can increase confidence which may create a positive association towards these type of interventions. We suggest measuring the level of feeling criticized per type of intervention instead of the general level of feeling criticized.

Third, we explored the functional preference of policemen concerning an RM in an application simulating preventive frisking. This is indisputably different from actually executing preventive frisking in real life. This aspect may impact the experience of interacting with the RM. Though, it is already argued that the RM could potentially be used as a training tool in an application, rather than include it in a real life operation [7].

Finally, we think that the achieved level of reflection in the participants may be a bit superficial and may be subject to confabulation [20]. Confabulation refers to the non-pathological phenomenon of providing reasons for actions that seem to describe deliberations while these reasons are not the causes of these actions. This does not imply we do not think our participants were trustworthy, it merely implies that we are not always aware of our own motivations because it is ultimately deeply complex because of our own extensive history of experience, actions and decisions. According to Staller and Koerner (2020), there are three levels of reflection: 'Am I doing things right?', 'Am I doing the right things?' and 'How do I decide if I am doing the right things?'. These three levels measure against a right/wrong template; assumptions in relation to actions; and assumptions in terms of perspectives, attitudes, etc respectively. They argue that is is possible to reflect without broadening perspectives if solely level 1 is included in the reflection. When inspecting the motivations of why (not) to frisk a particular individual, we reason that we have solely achieved the first level of reflection. We suggest further research into how to be able to achieve a deeper level of reflection.

Acknowledgements

I am thankful to my supervisor Pim Haselager, who supported me and kept me inspired throughout this project. I would like to thank Anaïs de Graaf for implementing the application for me and collaborating together in the beginning of this thesis. Furthermore, I am grateful for the policeman in training who engaged in the expert interview and all students and policemen who made time for participating in my experiments. I also genuinely appreciate the interest in my thesis others have shown.

Bibliography

[1] NOS, "Veelbesproken proef preventief fouilleren van start in amsterdam," 2021. https://nos.nl/artikel/2395993-veelbesproken-proef-preventief-fouilleren-van-start-in-amsterdam.

[2] "Artikel 151b." https://wetten.overheid.nl/jci1.3:c:BWBR0005416\&titeldeel=III\&hoofdstuk=IX\&artikel=151b\&z=2022-01-01\&g=2022-01-01.

[3] "Artikel 174b." https://wetten.overheid.nl/jci1.3:c:BWBR0005416&titeldeel=III&hoofdstuk=XI&artikel=174b&z=2022-01-01&g=2022-01-01.

[4] "Bonden: geen waarnemers bij preventief fouilleren." https://www.politiebond.nl/actueel/nieuws-and-blog/bonden-geen-waarnemers-bij-preventief-fouilleren/.

[5] N. Cornelissen, "A modular approach to increasing meaningful human control with decision support systems.," 2021.

[6] C. D. Guess, S. J. Donovan, D. Naslund, and C. D. Güss, "Improving dynamic decision making through training and self-improving dynamic decision making through training and self-reflection reflection improving dynamic decision making through training and self-reflection," 2015.

[7] A. de Graaf (submitted), "Reflection machines in a preventive frisking context: Useful applications and design decisions."

[8] D. A. Schön, "Educating the reflective practitioner: Toward a new design for teaching and learning in the professions. jossey-bass," 1987.

[9] C. R. Teal, R. E. Shada, A. C. Gill, B. M. Thompson, E. Frugé, G. B. Villarreal, and P. Haidet, "When best intentions aren't enough: helping medical students develop strategies for managing bias about patients.," *Journal of general internal medicine*, vol. 25 Suppl 2, pp. 115–118, 3 2010.

[10] S. Hatzipanagos and S. Lygo-Baker, "Teaching observations: Promoting development through critical reflection," *Journal of Further and Higher Education*, vol. 30, pp. 421–431, 2006.

[11] U. Pagallo, G. Sartor, T. Burri, F. S. D. Sio, and J. van den Hoven, "Meaningful human control over autonomous systems: a philosophical account," vol. 5, p. 28, 2018.

[12] A. Matthias, "The responsibility gap: Ascribing responsibility for the actions of learning automata."

[13] B. van der Stigchel, "Intelligent decision support in medical triage: Are people robust to biased advice?," submitted.

[14] S. Brayne, "Big data surveillance: The case of policing," *American Sociological Review*, vol. 82, pp. 977–1008, 10 2017.

[15] J. Ehrlinger, T. Gilovich, and L. Ross, "Peering into the bias blind spot: People's assessments of bias in themselves and others," *Personality and Social Psychology Bulletin*, vol. 31, pp. 680–692, 5 2005.

[16] W. Huisman, J. de Keijser, and F. Weerman, "Etnisch profileren door de politie," 2017. `doi:https://www.navigator.nl/document/id7351b226465d436884b8f6650a24359b/delikt-en-delinkwent-criminologie`.

[17] "Harde eisen." `https://kombijde.politie.nl/agent-worden/intake/harde-eisen`.

[18] N. van der Bijl, S. P. A. Bijvoet, P. C. van Dorst, S. Matthijssen, J. A. M. Pronk, M. Roeloffzen, and B. J. Vegter, "Waarborgen bij preventief fouilleren," 2011.

[19] C. T. Kydd, "Cognitive biases in the use of computer-based decision support systems," *OMEGA Int J of Mgmt Scl*, vol. 17, pp. 335–344, 1989.

[20] M. Slors, L. D. Bruin, and D. Strijbos, *Philosophy of mind, brain and behaviour*. Boom Amsterdam, 2015. p. 176.

Appendix A

Appendix

A.1 RM interventions

Interventions used in our application

case 01 -

case 02 "Consider that 95% of all weapon carriers of this population carry luggage."

case 03 -

case 04 "How certain are you about your choice to frisk this individual (1 = not certain at all, 5 = very certain)? Why?"

case 05 "Consider that 100% of all people you have decided to frisk carry luggage."

case 06 "Are you sure that your motivation leads to an optimal decision? Why (not)?"

case 07 "Consider that 70% of all weapon carriers of this population are female"

case 08 "Consider that 66,7% of all people you have decided to frisk are female"

case 09 "How certain are you about your choice to frisk this individual (1 = not certain at all, 5 = very certain)? Why?"

case 10 "Are you sure that your motivation leads to an optimal decision? Why (not)?"

case 11 "Are you sure you want to frisk this individual? Why (not)?"

case 12 "How certain are you about your choice to frisk this individual (1 = not certain at all, 5 = very certain)? Why?"

case 13 "Consider that 81% of all people of the population walking past this spot are in the age range of [20-40]."

case 14 "Consider that 50% of all people of the population walking past this spot are female"

case 15 "Consider that 70% of all people of the population walking past this spot owns a weapon"

case 16 "Consider that 30% of all people of the population walking past this spot are male"

case 17 "Consider that [X]% of all the people you decided to frisk up to now have exotic features."

case 18 "Consider that [X]% of all the people you decided to frisk up to now are men."

case 19 "Consider that [X]% of all the people you decided not to frisk up to now are in the age range of [20-40]."

case 20 "Consider that [X]% of all the people you decided to frisk up to now are men."

case 21 (only students) "Consider that 70% of all weapons carriers of this population is a woman"

case 22 (only students) "Are you sure about your decision to frisk this person? Why (not)?"

case 23 (only students) "Are you sure your focus will lead to an optimal decision?"

case 24 (only students) "Consider that 50% of this population carries a headscarf"

A.2 Forms for the participant

A.2.1 Consent Forms

Radboud Universiteit

INFORMATION LETTER

for participation in scientific research: Support in Preventive Frisking

1. The research

This study consists of four blocks. It involves an application in which you get to see a total of 28 videos of people walking on a square. Your job is to answer questions about some of these people with respect to preventive frisking *(in Dutch: preventief fouilleren)*. After the first three blocks you are asked to fill out a survey about your experience. More detailed information will be given during the experiment.

The current study has been reviewed by the Ethics Committee Social Sciences (ECSS).

2a. Use of your personal data

In this study we gather personal data about your experience with a supporting system in preventive frisking.

2b. Confidentiality of your data and data processing

The gathered information will be processed anonymously. This means that later results cannot be traced back to you. A consequence of this is that we cannot inform you about any personal results. In order to disguise your identity, only anonymized research data are to be used in reports and publications regarding the research.

The consent form signed by you will be kept for 1 year upon completion of the research. Your anonymized research data will be stored for 1 year after the research has been completed.

2d Sharing your data

Due to the importance of control, reuse and/or replication of research results, research data (including any anonymous personal data) are shared with or made available to other researchers. Your data will be anonymized prior to this form of sharing. This means that you can no longer be identified on the basis of these data. If you do not want your anonymized data to be shared, you can request to have your data deleted up to a maximum of 1 month after completion of the research.

2e Right of access by supervisory authorities to inspect the research's compliance with ruling guidelines

Some persons and organizations must have access to your personal and research data. This is necessary in order to test whether the research has been carried out properly and reliably. These persons and supervisory authorities inspecting your data for verification include: the supervisor in this thesis of the Radboud University: Pim Haselager. They are held to

1

inspecting your data on a strictly confidential basis. You will be asked to grant permission for this access. If you refuse to do so, you cannot participate in the study.

2f. Additional information on your rights regarding the processing of your personal data

Radboud University is responsible for compliance with the General Data Protection Regulation (GDPR) when processing your personal data. The researcher ensures that your privacy and the conditions attached to it are safeguarded and he/she adheres to the Dutch code of conduct for scientific integrity and university policy regarding the storage and management of personal and research data when conducting this research. You have the right to withdraw your consent for the processing of your personal data at any time. Your personal data will then be deleted. You can find the Radboud University Privacy Statement at: https://www.ru.nl/english/vaste-onderdelen/privacy-statement-radboud-university/. If you have any questions about your privacy, please contact the Privacy Officer Faculty of Social Sciences (P.Janssen@socsci.ru.nl). For general questions, please contact the office of the Data Protection Officer of Radboud University via privacy@ru.nl. More information about your rights in the processing of your personal data can be found at https://www.ru.nl/privacy/english/protection-personal-data/data-subjects-rights/ and on the website of the Dutch Data Protection Authority (https://autoriteitpersoonsgegevens.nl/en).

3. Voluntary participation

Your participation in this study is entirely voluntary. If you decide not to participate, there will be no consequences. If, during the course of the research, you wish to withdraw your consent and terminate your participation, you have every right to do so at all times. Again, there will be no adverse consequences for you.

4. Conclusion

If you have any questions, comments or concerns about this study, please contact the researcher responsible: Hannah Venhuizen.

With kind regards,

Hannah Venhuizen
Hannah.venhuizen@hotmail.com
Student Bachelor Artificial Intelligence, Radboud University

2

Radboud Universiteit

CONSENT FORM

for participation in scientific research: Support in Preventive Frisking

I herewith confirm that:
- I have been satisfactorily informed of the study in writing;
- I have read the written information
- I have been given the opportunity to ask questions about the study;
- my questions have been answered satisfactorily;
- I have been given ample opportunity to think carefully about participating in the study;
- I participate in the study entirely on a voluntary basis.

I understand that:
- I have the right to withdraw my consent at any time without having to state reasons and without fear of adverse consequences by contacting Hannah Venhuizen at hannah.venhuizen@hotmail.com;
- I have the right to have my research data deleted up until 1 month after completion of the research
- I have the right to withdraw my consent for the (further) processing of my personal data; my personal data are processed in accordance with the applicable European privacy regulations;
- my personal data are processed in accordance with the privacy statement of Radboud University (https://www.ru.nl/english/vaste-onderdelen/privacy-statement-radboud-university/);
- the tests and questionnaires used are not medical/clinical tests, but the researchers nonetheless have an obligation to inform me about scores that may be of personal clinical interest.

I agree that:
- my research data within this research will be obtained for scientific purposes and will be available for verification, reuse and replication for 1 year;
- the signed consent form with my personal data is kept for 1 year;
- my personal data, which are obtained for administrative purposes only, will be kept for a maximum of 1 month after completion of the research.
- supervisory authorities may inspect my personal and research data for the purpose of the research.

I agree to participate in the study.
Name:

Signature:

Date:

To be completed by the researcher:
I, the undersigned, herewith declare that the above-mentioned person has been informed both in writing and orally about the above-mentioned research.

Name:
Position/research institute:
Signature:
Date:

3

A.2.2 Information Forms

INFORMATION LETTER – Elaborate information 0.1a for participation in scientific research: Support in Preventive Frisking

Imagine a population with a very high crime rate. We call this population Z. In population Z, approximately 70 out of 100 people own a weapon and most of them are not scared to use it. Note, that the distribution of crime over the population may differ from your knowledge about people in general. The government wants to put the weapon rate down and decides to commence with a project in preventive frisking (in Dutch: preventief fouilleren).

This preventive frisking is done with the use of smart cameras. The cameras detect and analyse every individual. Then, an algorithm selects individuals based on inferences made from data of population Z. However, these algorithms are not perfect, thus it is essential to keep a human in charge. This is where you come in, imagine yourself -as the police- walking on the square and obtaining images of the camera with individuals pointed out. Your job is to make the decision whether you want to frisk this individual or not. In order to support you we added another system. This system poses reflective questions or notifications. The goal of this system is to increase the reflection process.

First you will see a fragment of the camera recording of an individual selected by the algorithm together with the question whether you want to frisk him/her. Then you may receive a reflective question or notification after which you get the chance to change your initial decision. The experiment consists of 3 blocks. After each block, you will be asked to fill out an evaluation form in the browser. You can do this by pressing the windows button on the keyboard and then navigate to the chrome browser.

If you have any questions during the experiment or are not sure what to do, do not hesitate to ask. Now, please notify the experimenter that you have read this.

INFORMATION LETTER – Elaborate information 0.1b for participation in scientific research: Support in Preventive Frisking

Imagine a population with a very high crime rate. We call this population Z. In population Z, approximately 70 out of 100 people own a weapon and most of them are not scared to use it. Note, that the distribution of crime over the population may differ from your knowledge about people in general. The government wants to put the weapon rate down and decides to commence with a project in preventive frisking.

This is where you come in, imagine yourself -as the police- walking on the square. Your job is to make the decision whether you want to frisk some individual or not. In order to support you we added a system. This system poses reflective questions or notifications. The goal of this system is to increase the reflection process.

First you will see a fragment of the camera recording of an individual pointed out randomly with the question whether you want to frisk him/her. Then you may receive a reflective question or notification after which you get the chance to change your initial decision. The experiment consists of 3 blocks. After each block you will be asked to fill out an evaluation form in the browser. You can do this by pressing the windows button on the keyboard and then navigate to the chrome browser.

If you have any questions during the experiment or are not sure what to do, do not hesitate to ask. Now, please notify the experimenter that you have read this.

A.2.3 Debriefing forms

INFORMATION LETTER – debriefing – a_en
for participation in scientific research: Support in Preventive Frisking

Dear participant,

First, thank you for completing the experiment.

Second, we want to tell you the actual goal of the experiment. We have conducted this experiment in order to evaluate the experience of a potential Reflection Machine. This (still non-existing) system poses reflective questions or notifications to make you reconsider your decisions. The goal of this is to increase meaningful human control.

The so-called algorithm we told you about does <u>not</u> exist and is <u>not</u> the goal of this research. No system was involved in making the selection of individuals. However, the algorithm story is included in this experiment to test whether the experience of making use of such an algorithm would influence the experience of the use of a Reflection Machine. Half of the participants is told about the algorithm, whereas half is not. In this way, we can analyze whether it makes a difference in the user experience.

We also analyze your experience with other variables like type of questions, as you may have noticed in the surveys.

We would like to stress again that you can contact the researcher responsible if you have any questions, comments or concerns about this study.

INFORMATION LETTER – debriefing – a_nl
for participation in scientific research: Support in Preventive Frisking

Beste deelnemer,

Allereerst bedankt voor het voltooien van het experiment.

Ten tweede willen we u nu het eigenlijke doel van het experiment vertellen. We hebben dit experiment uitgevoerd om de ervaring van een potentiële Reflection Machine te evalueren. Dit (nog niet bestaande) systeem stelt reflectieve vragen of notificaties om je beslissingen te laten heroverwegen. Het doel hiervan is om zinvolle menselijke controle te vergroten.

Het zogenaamde algoritme waar we u over vertelden bestaat <u>niet</u> en is ook <u>niet</u> het doel van dit onderzoek. Er was geen systeem betrokken bij het maken van de selectie van individuen. Het verhaal van het algoritme is echter in dit experiment opgenomen om te testen of de ervaring van het gebruik van zo'n algoritme de ervaring van het gebruik van een Reflection Machine zou beïnvloeden. De helft van de deelnemers wordt verteld over het algoritme, de andere helft niet. Op deze manier kunnen we analyseren of het een verschil maakt in de gebruikerservaring.

We analyseren ook uw ervaring met andere variabelen, zoals het type vragen, zoals u wellicht heeft opgemerkt in de enquêtes.

We willen nogmaals benadrukken dat u met vragen, opmerkingen of zorgen over dit onderzoek contact kunt opnemen met de verantwoordelijke onderzoeker.

Bedankt!

INFORMATION LETTER – debriefing – b_en
for participation in scientific research: Support in Preventive Frisking

Dear participant,

First of all, thank you for completing the experiment.

Second, we now want to tell you the actual purpose of the experiment. We conducted this experiment to evaluate the experience of a potential Reflection Machine. This (not yet existing) system asks reflective questions or notifications to make you reconsider your decisions. The purpose of this is to stimulate awareness of underlying motivations.

This experiment consists of two groups. Half of the participants are told about an algorithm, the other half are not. You belong to the second group. The others will read the following in the information form prior to the experiment:

"This preventive frisking is done with the use of smart cameras. The cameras detect and analyse every individual. Then, an algorithm selects individuals based on inferences made from data of population Z. However, these algorithms are not perfect, thus it is essential to keep a human in charge. This is where you come in, imagine yourself -as the police- walking on the square and obtaining images of the camera with individuals pointed out. Your job is to make the decision whether you want to frisk this individual or not."

The so-called algorithm that we told you about does <u>not</u> exist and is <u>not</u> the aim of this study. However, the story of the algorithm is included in this experiment to test whether the experience of using such an algorithm would influence the experience of using a Reflection Machine. In this way we can analyze whether it makes a difference in the user experience.

We also analyze your experience with other variables, such as the type of questions, as you may have noticed in the surveys.

We would like to emphasize once again that you can contact the responsible researcher with any questions, comments or concerns about this study.

Thank you!

INFORMATION LETTER – debriefing – b_nl
for participation in scientific research: Support in Preventive Frisking

Beste deelnemer,

Allereerst bedankt voor het voltooien van het experiment.

Ten tweede willen we u nu het eigenlijke doel van het experiment vertellen. We hebben dit experiment uitgevoerd om de ervaring van een potentiële Reflection Machine te evalueren. Dit (nog niet bestaande) systeem stelt reflectieve vragen of notificaties om je beslissingen te laten heroverwegen. Het doel hiervan is om de bewustwording van onderliggende beweegredenen te stimuleren.

Dit experiment bestaat uit twee groepen. De helft van de deelnemers wordt verteld over een algoritme, de andere helft niet. U behoort tot de tweede groep. De anderen krijgen het volgende te lezen in het informatie formulier voorafgaand aan het experiment:

"Dit preventief fouilleren wordt ondersteund door het gebruik van slimme camera's. De camera's detecteren en analyseren elk individu. Vervolgens maakt een algoritme een selectie individuen op basis van gegevens van populatie Z. Deze algoritmen zijn echter niet perfect, dus het is essentieel om een mens aan het roer te houden. Dit is waar u een rol speelt, stelt u zich voor dat u in functie op het plein loopt en beelden van de camera krijgt waarop individuen worden uitgelicht. Het is uw taak om de beslissing te nemen of u deze persoon wilt fouilleren of niet."

Het zogenaamde algoritme waar we u over vertelden bestaat dus <u>niet</u> en is ook <u>niet</u> het doel van dit onderzoek. Het verhaal van het algoritme is echter in dit experiment opgenomen om te testen of de ervaring van het gebruik van zo'n algoritme de ervaring van het gebruik van een Reflection Machine zou beïnvloeden. Op deze manier kunnen we analyseren of het een verschil maakt in de gebruikerservaring.

We analyseren ook uw ervaring met andere variabelen, zoals het type vragen, zoals u wellicht heeft opgemerkt in de enquêtes.

We willen nogmaals benadrukken dat u met vragen, opmerkingen of zorgen over dit onderzoek contact kunt opnemen met de verantwoordelijke onderzoeker.

Bedankt!

A.3 Survey

A.3.1 Statements

1. "I think the selection made by the algorithm is very accurate"
2. "After a reflective question/notification I actively reconsider my decision about the frisking (in Dutch: fouilleren)"
3. "I would have picked the same individuals as the algorithm"
4. "How do you experience the amount of reflective questions and notifications?"
5. "I think the concept of the algorithm could be helpful in the field of preventive frisking"
6. "How do you experience the quality of the reflective questions and notifications?"
7. "I felt criticized by the reflective questions and notifications"
8. "The reflective question or notification helped me think about potential bias (in Dutch: vooringenomen standpunten)"
9. "You saw the following questions during the experiment. Please sort them in order of relevance (with the top one being the most relevant)" [1]

   C1 "Are you sure you want to frisk this individual? Why (not)?"

   C2 "How certain are you about your choice to frisk this individual (1 = not certain at all, 5 = very certain)? Why?"

   C3 "Are you sure that your motivation leads to an optimal decision? Why (not)?"

   P1 "Consider that [X] % of all people of the population walking past this spot [...]"

   P2 "Do you think this person deviates from the population in a particular aspect? Why (not)?"

   P3 "Consider that [X] % of all weapons carriers of this population [...]"

   H1 "Up to now you have decided (not) to frisk [X]% of all the people presented to you"

   H2 "Consider that [X]% of all the people you decided (not) to frisk up to now [...]

[1] Note that the labels are not shown to the participant.

A.3.2 Data

The data can be found in the GitHub repository (https://github.com/Hann-v0/ReflectionMachine.git)

A.4 Screenshots flow application

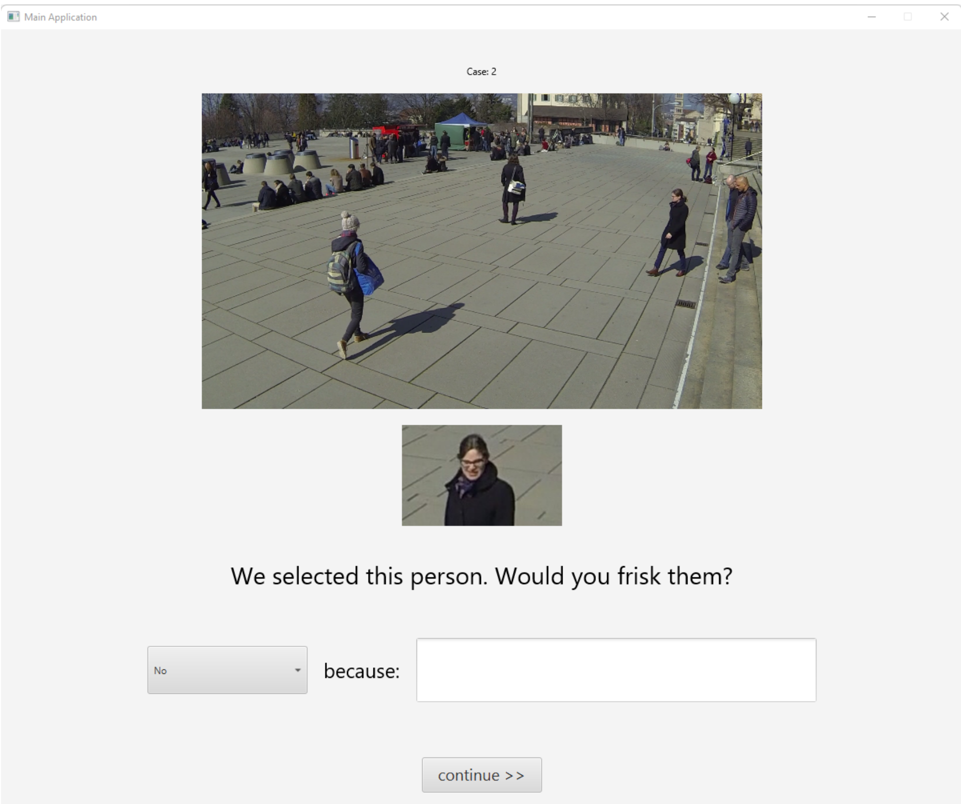


Figure A.1: Frame 1

This frame includes a video clip and highlighted individual while asking if the user wants to frisk this person

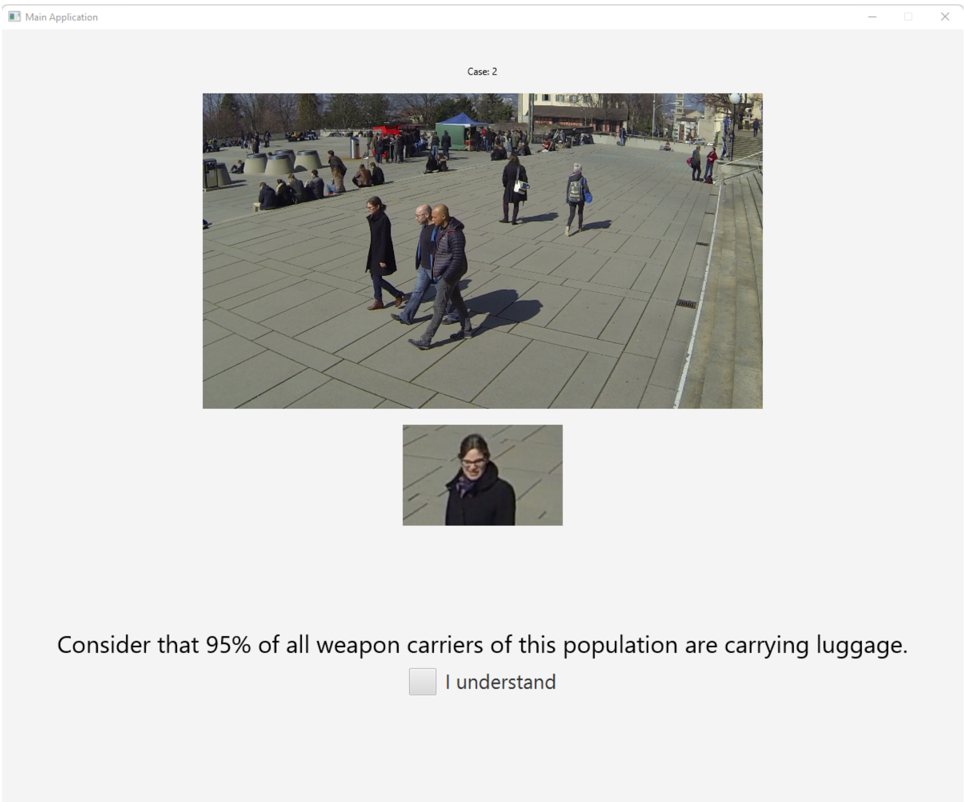


Figure A.2: Frame 2

This frame presents a reflective intervention to the user

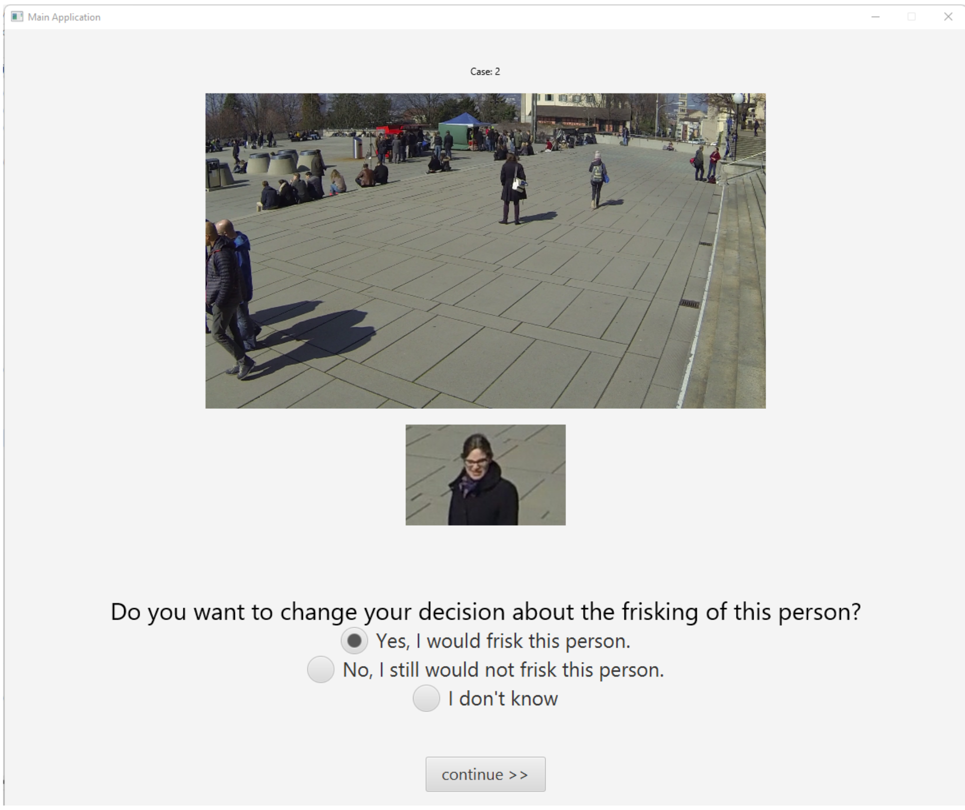


Figure A.3: Frame 3

This frame gives the user the opportunity to change the initial decision