Information sheet. ***Please read the following text carefully before starting.***

You are being invited to participate in a research study titled *Explainable AI for human supervision over firefighting robots*. This study is being executed by Ruben Verhagen, Elena Negrila, Bogdan Pietroianu, Dafni Pandeva, Yi Wu, and Elena Ibanez, as part of the course *Research Project* at Delft University of Technology.

Humans and AI systems increasingly work together when solving problems or executing tasks. In such human-AI teams, higher levels of AI autonomy should be combined with meaningful ways of human control. One approach is to let a human make all moral decisions, while the AI system can make non-moral decisions. In this study, you will be asked to collaborate with an autonomous AI system that allocates moral decision making to you while making non-moral decisions itself. Your task will be to search and rescue victims during a simulated 2D firefighting task. Your goal will not be to rescue as many victims as fast as possible, but rather to achieve the best outcomes for both victims and firefighters. For example, by minimizing the loss of victims while maximizing firefighter safety.

The purpose of this research study is to gain insights into how you evaluate the collaboration with the AI system. We will collect this data with a few questionnaires after the task. We expect the study to take approximately 30 minutes to complete, and the results will be used for research purposes only. Your participation in this study is entirely voluntary and you can withdraw at any time, without having to give a reason.

The experiment may increase your workload since it involves using your keyboard and mouse to navigate and communicate, while also receiving messages and trying to complete the task. However, we took several steps to ensure the experimental design is user-friendly. In addition, we will give you time to get familiar with the simulated environment, controls, and messaging system.

We believe there are no known risks associated with this research study; however, as with any online related activity the risk of a breach is possible. To the best of our ability, your answers in this study will remain confidential. We will minimize any risks by only asking you about your gender, age range, level of education, trait measures, and gaming experience. This will make re-identification impossible (i.e., your data is anonymous). Since your data will be anonymous, you cannot request your data to be removed after completion of the study. We will archive your anonymized data at 4TU.ResearchData for at least 10 years, so it can be used for future research and learning. This data will be publicly available for non-commercial use only.

In case of questions or complaints, you can contact **R.S.Verhagen@tudelft.nl**.

Introducing Consent. It is a good research practice to have participants provide "informed written consent". For the following statements, please tick the appropriate boxes:

Consent1. I have read and understood the study information, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.

- (●) Yes
- ( ) No

Consent2. I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and withdraw from the study at any time, without having to give a reason.

- (●) Yes
- ( ) No

Consent3. I understand that taking part in the study involves fully completing the task and surveys that I will be presented with.

- (●) Yes
- ( ) No

Consent4. I understand and agree that answers I provide will be anonymized and be publicly available for non-commercial use only.

- (●) Yes
- ( ) No

Consent5. I understand that I cannot request for the information that I provide to be removed from storage after completing the surveys.

- (●) Yes
- ( ) No

Consent6. I give permission for answers and information that I provide to be anonymized and archived at 4TU.ResearchData so it can be used for future research and learning.

- (●) Yes
- ( ) No

Consent7. I agree to take part in this study.

- (●) Yes
- ( ) No

Gender. What gender do you identify as?

- (●) Female
- ( ) Male
- ( ) Other [ ]
- ( ) Prefer not to say

Age. What is your age?

- (●) 18 - 24 years old
- ( ) 25 - 34 years old
- ( ) 35 - 44 years old
- ( ) 45 - 54 years old
- ( ) 55 - 64 years old
- ( ) 65+ years old
- ( ) Prefer not to say

Education. What is the highest degree or level of education you have completed?

- ( ) No schooling completed
- ( ) Some high school, no diploma
- ( ) High school graduate
- (●) Some college credit, no degree
- ( ) Associate degree
- ( ) Bachelor's degree
- ( ) Master's degree
- ( ) Ph.D. degree or higher
- ( ) Prefer not to say

Gaming. How much video/computer gaming experience do you have?

- ( ) None at all
- (●) A little
- ( ) A moderate amount
- ( ) A considerable amount
- ( ) A lot

Risk propensity 1. Please indicate the degree to which you agree/disagree with the following statements.

	totally disagree								totally agree
	1	2	3	4	5	6	7	8	9
Safety first	○	○	○	○	○	○	◉	○	○
I do not take risks with my health	○	○	○	○	○	◉	○	○	○
I prefer to avoid risks	○	○	○	○	○	◉	○	○	○
I take risks regularly	○	○	◉	○	○	○	○	○	○
I really dislike not knowing what is going to happen	○	○	○	○	○	○	◉	○	○
I usually view risks as a challenge	○	○	○	○	○	◉	○	○	○

Risk propensity 2.

	risk avoider								risk seeker
	1	2	3	4	5	6	7	8	9
I view myself as a ...	○	○	○	◉	○	○	○	○	○

Introduction. Please indicate the degree to which you agree/disagree with the following statements.

Q2. I usually trust technology until there is a reason not to.

strongly disagree	somewhat disagree	neither agree nor disagree	somewhat agree	strongly agree
○	○	○	◉	○

Q3. For the most part, I distrust technology.

strongly disagree	somewhat disagree	neither agree nor disagree	somewhat agree	strongly agree
○	◉	○	○	○

Q4. In general, I would rely on technology to assist me.

strongly disagree	somewhat disagree	neither agree nor disagree	somewhat agree	strongly agree
○	○	○	◉	○

Q5. My tendency to trust technology is high.

strongly disagree	somewhat disagree	neither agree nor disagree	somewhat agree	strongly agree
○	○	○	◉	○

Q6. It is easy for me to trust technology to do its job.

strongly disagree	somewhat disagree	neither agree nor disagree	somewhat agree	strongly agree
○	○	○	◉	○

Q7. I am likely to trust technology even when I have little knowledge about it.

strongly disagree	somewhat disagree	neither agree nor disagree	somewhat agree	strongly agree
○	○	◉	○	○

Introduction. The following questions will ask you about your utilitarian ethical beliefs and values. Please indicate the degree to which you agree/disagree with the following statements.

Sacrifice. If the only way to save another person's life during an emergency is to sacrifice one's own leg, then one is morally required to make this sacrifice.

strongly disagree	somewhat disagree	neither agree nor disagree	somewhat agree	strongly agree
○	◉	○	○	○

Harm. It is morally right to harm an innocent person if harming them is a necessary means to helping several other innocent people.

strongly disagree	somewhat disagree	neither agree nor disagree	somewhat agree	strongly agree
○	○	○	◉	○

Donate. From a moral point of view, we should feel obliged to give one of our kidneys to a person with kidney failure since we don't need two kidneys to survive, but really only one to be healthy.

strongly disagree	somewhat disagree	neither agree nor disagree	somewhat agree	strongly agree
○	◉	○	○	○

Oppression. If the only way to ensure the overall well-being and happiness of the people is through the use of political oppression for a short, limited period, then political oppression should be used.

strongly disagree	somewhat disagree	neither agree nor disagree	somewhat agree	strongly agree
○	◉	○	○	○

Well-being. From a moral perspective, people should care about the well-being of all human beings on the planet equally; they should not favor the well-being of people who are especially close to them either

physically or emotionally.

strongly disagree	somewhat disagree	neither agree nor disagree	somewhat agree	strongly agree
○	◉	○	○	○

Torture. It is permissible to torture an innocent person if this would be necessary to provide information to prevent a bomb going off that would kill hundreds of people.

strongly disagree	somewhat disagree	neither agree nor disagree	somewhat agree	strongly agree
○	◉	○	○	○

Help. It is just as wrong to fail to help someone as it to actively harm them yourself.

strongly disagree	somewhat disagree	neither agree nor disagree	somewhat agree	strongly agree
○	○	○	◉	○

Collateral. Sometimes it is morally necessary for innocent people to die as collateral damage - if more people are saved overall.

strongly disagree	somewhat disagree	neither agree nor disagree	somewhat agree	strongly agree
○	○	○	◉	○

Money. It is morally wrong to keep money that one doesn't really need if one can donate it to causes that provide effective help to those who will benefit a great deal.

strongly disagree	somewhat disagree	neither agree nor disagree	somewhat agree	strongly agree
○	◉	○	○	○

Version. Please specify your experiment version.

- ○ 1
- ○ 2
- ◉ 3
- ○ 4
- ○ 5
- ○ 6

Introduction.
We will now give you time to get familiar with the environment, controls, and messaging system. After that, we will start the official task. Please ask the experimenter to start the tutorial.

Introduction. The following questions will ask you about your perception of Brutus during completion of the task.

MDMT. Please rate Brutus using the scale from 0 (not at all) to 7 (very). If a particular item does not seem to fit Brutus in the situation, please select the option that says "does not fit".

	not at all 0	1	2	3	4	5	6	very 7	does not fit
Reliable	○	○	○	○	○	○	◉	○	○
Sincere	○	○	○	○	○	◉	○	○	○
Capable	○	○	○	○	○	○	◉	○	○
Ethical	○	○	○	○	○	◉	○	○	○
Predictable	○	○	○	○	○	◉	○	○	○
Genuine	○	○	○	○	○	◉	○	○	○
Skilled	○	○	○	○	○	○	◉	○	○
Respectable	○	○	○	○	○	◉	○	○	○
	not at all 0	1	2	3	4	5	6	very 7	does not fit
Someone you can count on	○	○	○	○	○	○	◉	○	○
Candid (i.e., marked by honest sincere expression)	○	○	○	○	○	◉	○	○	○
Competent	○	○	○	○	○	○	◉	○	○
Principled	○	○	○	○	○	◉	○	○	○
Consistent	○	○	○	○	○	○	◉	○	○
Authentic	○	○	○	○	○	○	◉	○	○
Meticulous (i.e., marked by great attention to detail)	○	○	○	○	○	○	◉	○	○
Has integrity	○	○	○	○	○	○	◉	○	○
	not at all 0	1	2	3	4	5	6	very 7	does not fit

Introduction. The following questions will ask you about your satisfaction with the explanations provided by Brutus **when it allocated decision-making to you or itself**.

Understanding. From the explanations, I understand how Brutus works.

I disagree strongly	I disagree somewhat	I am neutral about it	I agree somewhat	I agree strongly
○	○	○	◉	○

Satisfaction. The explanations provided by Brutus are satisfying.

I disagree strongly	I disagree somewhat	I am neutral about it	I agree somewhat	I agree strongly
○	○	○	◉	○

Sufficiency. The explanations provided by Brutus have sufficient detail.

I disagree strongly	I disagree somewhat	I am neutral about it	I agree somewhat	I agree strongly
○	○	○	◉	○

Completeness. The explanations provided by Brutus seem complete.

I disagree strongly	I disagree somewhat	I am neutral about it	I agree somewhat	I agree strongly
○	○	○	◉	○

Use. The explanations provided by Brutus tell me how to use it.

I disagree strongly	I disagree somewhat	I am neutral about it	I agree somewhat	I agree strongly
○	○	◉	○	○

Usefulness. The explanations provided by Brutus are useful to my goals.

I disagree strongly	I disagree somewhat	I am neutral about it	I agree somewhat	I agree strongly
○	○	○	◉	○

Accuracy. The explanations provided by Brutus show me how accurate Brutus is.

I disagree strongly	I disagree somewhat	I am neutral about it	I agree somewhat	I agree strongly
○	○	○	◉	○

Trust. The explanations provided by Brutus let me judge when I should trust and not trust Brutus.

I disagree strongly	I disagree somewhat	I am neutral about it	I agree somewhat	I agree strongly
○	○	○	◉	○

Location Data

Location: (52.3787, 4.7941)

Source: GeoIP Estimation