Article Identifier
DOI: 10.22032/dbt.58928
URN: urn:nbn:de:gbv:ilm1-2023isc-127:3

60th ILMENAU SCIENTIFIC COLLOQUIUM
Technische Universität Ilmenau, 4 – 8 September 2023
URN: urn:nbn:de:gbv:ilm1-2023isc:1

Human uncertainty in interaction with a machine: establishing a reference dataset

Anne Rother (1), Gunther Notni (2), Alexander Hasse (3), Benjamin Noack (1), Christian Beyer (1), Jan Reißmann (3), Chen Zhang (2), Marco Ragni (3), Julia Arlinghaus (1), Myra Spiliopoulou (1)

(1) Otto-von-Guericke-University Magdeburg
(2) Technische Universität Ilmenau
(3) Technical University Chemnitz

ABSTRACT

We investigate the task of malformed object classification in an industrial setting, where the term ‘malformed’ encompasses objects that are misshapen, distorted, corroded or broken. Recognizing whether such an object can be repaired, taken apart so that its components can be used otherwise, or dispatched for recycling, is a difficult classification task. Despite the progress of artificial intelligence for the classification of objects based on images, the classification of malformed objects still demands human involvement, because each such object is unique. Ideally, the intelligent machine should demand expert support only when it is uncertain about the class. But what if the human is also uncertain? Such a case must be recognized before being dealt with. Goal of this research thread is to establish a reference dataset on human uncertainty for such a classification problem and to derive indicators of uncertainty from sensory inputs. To this purpose, we designed an experiment for an object classification scenario where the uncertainty can be directly linked to the difficulty of labelling each object. By thus controlling uncertainty, we intend to build up a reference dataset and investigate how different sensory inputs can serve as uncertainty indicators for these data.

Index Terms – Object classification, human annotator experiments, uncertainty in human annotations, uncertainty indicators in human experiments

1. INTRODUCTION

In the interaction between human and machine, the term "symbiosis" is becoming increasingly important. Several dimensions are considered here: Task, interaction, performance and experience [1]. In their Nature paper on cooperative intelligence [2], Dafoe et al. point out that ‘many deployed machine-learning models are trained either on massive data sets or in simulated environments that can generate years of experience in seconds […] By contrast, humans produce data slowly …’. Particular attention with respect to data sparsity must be given to the fact that there are limits on the number of sensors human workers can or want to carry on their bodies for the sake of data production. Hence, in all endeavours on understanding human behaviour, the demand for sensory inputs, especially those attached to the body, should be kept low. In this study, we investigate ways of capturing human uncertainty for an object classification task in an industrial setting. As uncertainty indicators we consider sensory inputs that can be captured with a smartphone, namely electrodermal activity (EDA) and electrocardial signal (ECG), next to externally measurable indicators, such as task duration. Our ultimate goal is to minimise the demand from on-body sensory inputs, promoting signal forms that are generated remotely from the human body.

The task we investigate is the classification of ‘malformed’ objects, i.e. objects that disagree with the production norm – either because of faults in the production machines or because the objects have been delivered to be re-factored or to be taken apart for recycling. The classification objective concerns a skewed class distribution, e.g. most objects cannot be re-factored (dominant class) while some of them have re-usable materials or components (non-dominant classes with one or more types of rare labels).

Nowadays, object classification tasks for physical objects produced or assembled in a production floor are expected to be performed by a model induced by a deep learning algorithm. The main challenge for this is that the algorithm demands an adequate number of examples per class for training; if all objects are unique, this demand is difficult to satisfy. An alternative is to classify the objects in an interactive scenario, where the model delivers to the human those objects that it cannot label with high certainty. In both the case where the human(s) must deliver all object labels for training, and in the case where the human(s) must label the objects that cannot be labelled by the machine, it is essential to assess the certainty of the human about the label(s) s/he assigns. Hence, we formulated following research question:

RQ1: How to identify robust indicators of human uncertainty?

To address this RQ, we establish an experimental setting where human uncertainty is controlled through (a priori known) difficulty of each object classification task, we measure sensory inputs of the experiment participants as well as unobtrusive difficulty indicators (such as task duration), and we study to what extend these sensory inputs and indicators predict difficulty. We aim to build up a reference dataset for measuring human uncertainty, and to report on the potential of different uncertainty indicators.

In our study [3], human annotators performed pairwise comparisons tasks on triplets that consisted of 10-dimensional medical instances from a cohort; duration of task completion, electrodermal activity and disagreement were used as difficulty indicators. We found that for some instances proximity across certain dimensions was misleading, in the sense that annotators consistently decided that a pair of instances inside a triplet were more similar than they truly were.

Research on uncertainty typically distinguishes between ‘epistemic uncertainty‘ which is due to inadequate information, and ‘aleatoric uncertainty’ which reflects the stochastic nature of the observations [4, 5]. Valdenegro-Toro and Mori have shown that, unexpectedly, epistemic and aleatoric uncertainty influence each other [5], while Ghandeharioun et al. identified a relationship between aleatoric uncertainty and annotator disagreement [4]. Epistemic uncertainty can be reduced through additional information, whereby one has to take the cost of information acquisition into account. In [6], we proposed a mechanism for cost-aware acquisition of features to reduce missingness and improve of a model learned over a stream of observations. In the context of our experiment, epistemic uncertainty could be reduced by using a fine calliper instrument, a magnification glass or both; such instruments would be needed for some measurements but not for others.

2. EXPERIMENT DESIGN

This experiment is a follow-up of our work on experimentally assessing the difficulty of annotating medical data in crowd working [3] [1].

[1] Approved under Institutional Review Board Certificate No. NI4HLBHn.

2.1 Experiment workflow

The experiment consists of n trials, conducted one after the other within one session. In each trial, the participant is asked to assign a metal cylinder to one of two classes, where the class assignment depends on whether the cylinder's diameter is within a given boundary (class 1) or not (class 0). The diameter measurement must be estimated by heart, but there is also a calliper that can be used for a rough estimation. At the end of each trial, the participant is also asked to state his/her certainty in the class assignment s/he made. The whole experiment session is expected to vary between 60 and 90 minutes.

An introductory session with instructions on the classification task and a training session on calliper usage will precede the experiment. The configuration parameters are:

- the number of cylinders n (same as the number of trials, upper bounded by 150);
- the number of 'easy' cylinders (the diameter can be easily estimated by heart) vs 'difficult' ones (the diameter is difficult to estimate by heart);
- the number of Likert-scale values that the participants can choose from to describe their uncertainty; we anticipate either three values (low, medium, high) or five ones.

Following technologies will be used in the experiment: sensor technology EdaMove 4 (movisens GmbH) for the measurement of electro-dermal activity (Galvanic skin response) and EcgMove 4 (movisens GmbH) for the measurement of ECG with adjoint Data Analyzer software for the analysis of both signals.

In Figure 1 the cylinders are shown, which the participants should annotate one after the other. These 'specimens' have an outer diameter between d and d + 0.1 mm, varied steps of 0.01 mm (tolerances are Gaussian distributed)

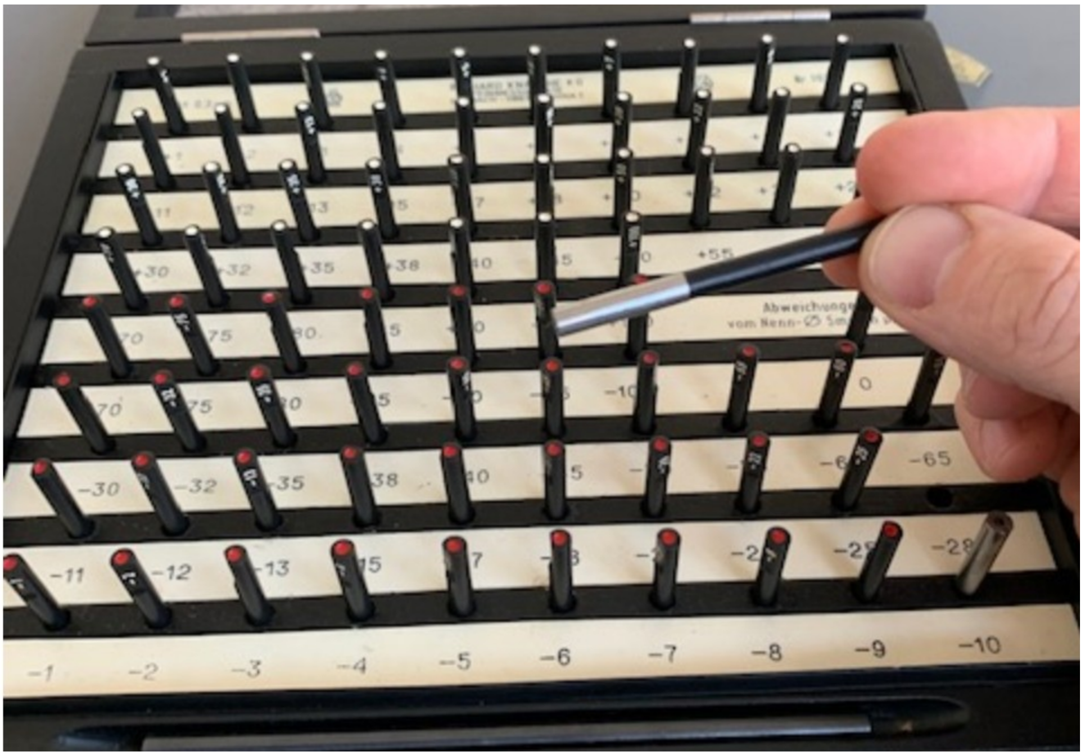
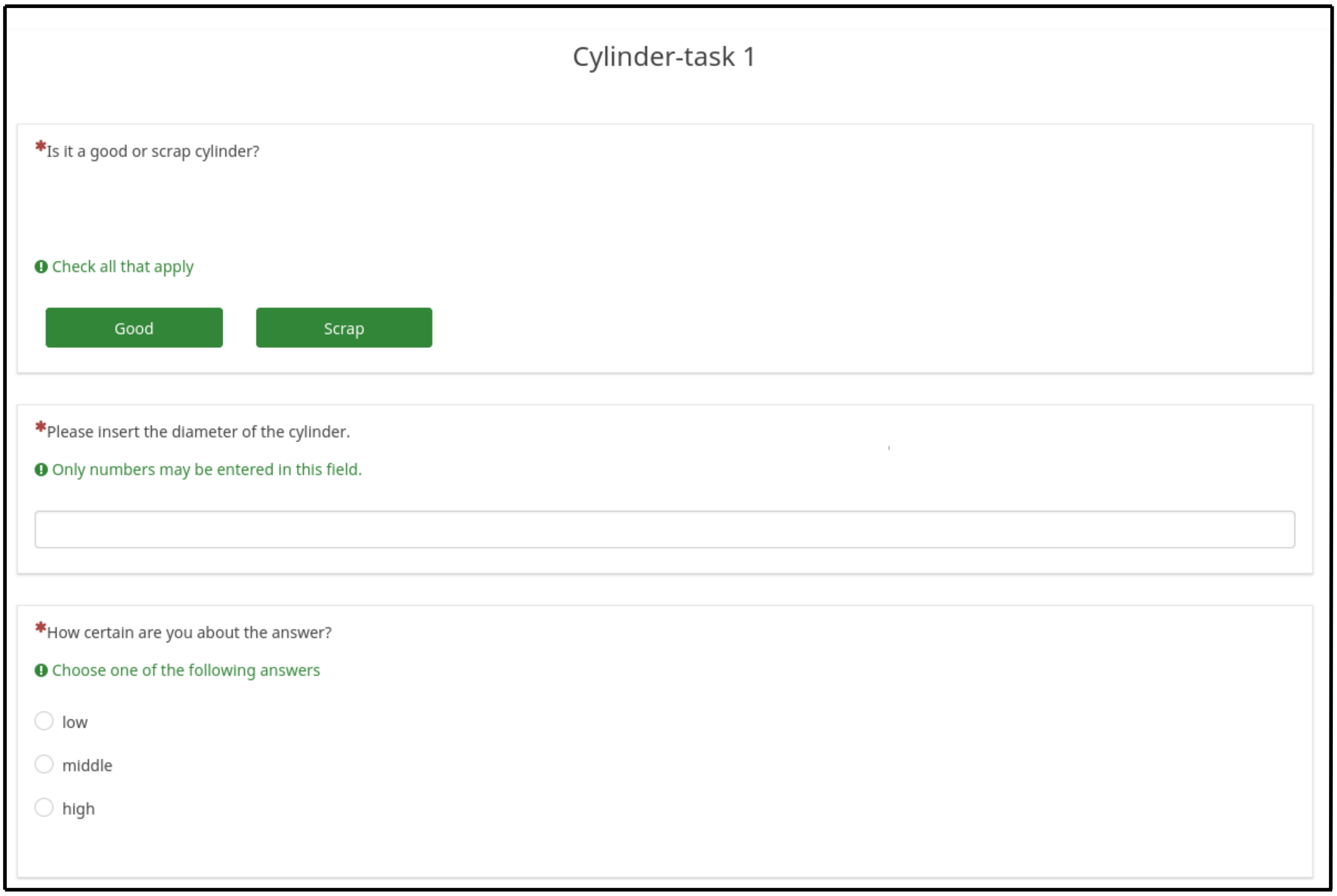


Fig 1. Overview of cylinder with different degrees of difficulty

For the annotation task, ‘good’ specimen have its outside diameter is up to d + 0.05 mm, and ‘scrap’ specimen: (i.e. metallic junk) have its outside diameter is higher.



Cylinder-task 1

*Is it a good or scrap cylinder?

Check all that apply

*Please insert the diameter of the cylinder.

Only numbers may be entered in this field.

*How certain are you about the answer?

Choose one of the following answers

○ low

○ middle

○ high

Fig. 2: Example of an annotation task

2.2 Inclusion and exclusion criteria

We will use stratified sampling to enforce gender balance and homogeneity with respect to age; volunteers will be recruited among students of the Otto-von-Guericke-University Magdeburg. To ensure broad coverage, we will also recruit students from international degrees. Therefore, the text of the questionnaire will be in English. The language of the introduction session can be German or English and can be decided on demand.

We will exclude participants who report impairments of sight and/or haptics that may prevent the proper use of the calliper.

2.3 Data to be recorded

We will record the time needed to complete each trial, the responses of the participants, the electrodermal signal from the palm of the non-dominant hand, ECG signal, and possibly respiration recording. For each participant, we will store age, gender (as stated by the participant), handedness (because it may affect the use of the calliper) and eyesight (because it may affect classification quality).

At the end of the experiment, each participant will fill a questionnaire with questions about the experiment, the perceived task difficulty and perceived fatigue, as well as sociodemographic characteristics for further information analyses.

3. ANTICIPATED RESULTS

We will investigate to what extent the following variables are predictive of correct label assignment ('correctness'): trial duration, uncertainty-as-stated, level of electrodermal activity, level of ECG signal and of respiratory signal. The levels will be derived through binning. Then, we will investigate the relationship between difficulty of each trial (i.e. of the cylinder labelled at each trial) and the values of the aforementioned variables.

We will further study the correlation between sociodemographics like gender and age to the predictive variables. We will vary the number of cylinders n and, for large n, we will devise mechanisms to capture fatigue, so that we can investigate the effects of fatigue upon correctness and upon uncertainty-as-stated.

REFERENCES

[1] J. Inga, M. Ruess, J.H. Robens, T. Nelius, S. Kille, P. Dahlinger, R. Thomaschke, G. Neumann, S. Matthiesen, S. Hohmann, A. Kiesel, "Human-machine symbiosis: A multivariate perspective for physically coupled human-machine systems", *International Journal of Human-Computer Studies*, *170*, 102926, February 2023.

[2] A. Dafoe, Y. Bachrach, G. Hadfield, E. Horvitz, K. Larson, T. Graepel, "Cooperative AI: machines must learn to find common ground", Nature, 593(7857), 33-36, 2021

[3] A. Rother, U. Niemann, T. Hielscher, H. Völzke, T. Ittermann, and M. Spiliopoulou, "Assessing the difficulty of annotating medical data in crowdworking with help of experiments", PLOS ONE, (16)7:1-26, Public Library of Science, July 2021.

[4] A. Ghandeharioun, B. Eoff, B. Jou, R.W. Picard, "Characterizing sources of uncertainty to proxy calibration and disambiguate annotator and data bias", IEEE/CVF International Conference on Computer Vision Workshop (ICCVW) , 42024206, IEEE, 2019.

[5] M. Valdenegro-Toro, D. S. Mori, "A deeper look into aleatoric and epistemic uncertainty disentanglement", IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 1508-1516), IEEE, 2022.

[6] M. Büttner, C. Beyer, M. Spiliopoulou "Reducing missingness in a stream through cost-aware active feature acquisition" IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2022.

CONTACTS

Anne Rother	email: anne.rother@ovgu.de ORCID: 0000-0002-6768-5871
Gunther Notni	email: gunther.notni@tu-ilmenau.de ORCID: 0000-0001-7532-1560
Alexander Hasse	email: alexander.hasse@mb.tu-chemnitz.de ORCID: 0000-0002-6737-5152
Benjamin Noack	email: benjamin.noack@ovgu.de ORCID: 0000-0001-8996-5738

Christian Beyer
email: christian.beyer@ovgu.de
ORCID: 0000-0001-8604-9523

Jan Reißmann
email: jan.reissmann@mb.tu-chemnitz.de
ORCID: 0009-0000-2825-9725

Cheng Zhang
email: chen.zhang@tu-ilmenau.de
ORCID: 0000-0001-8942-4624

Marco Ragni
email: marco.ragni@hsw.tu-chemnitz.de
ORCID: 0000-0003-2661-2470

Julia Arlinghaus
email: julia.arlinghaus@ovgu.de
ORCID: 0000-0002-0287-2086

Myra Spiliopoulou
email: myra@ovgu.de
ORCID: 0000-0002-1828-5759