Calibrating Trust Towards An Autonomous Image Classifier Dataset Overview:

Each row represents one observation (159) per participant (74) from our human-computer interaction experiment, where participants worked with an autonomous image classifier to classify images. ***Please Note*** that observation data was corrupted for one of our image stimuli (Unique_Image_Number: 72), and is therefore unavailable in this dataset. As a result, each participant is missing 1 trial from their data (74 missing observations), the location of which was randomly located within the 160 trials that each participant completed. An explanation of each variable in the dataset is provided below, and if you have any questions, please feel free to contact the authors:

Martin Ingram: m.ingram.1@research.gla.ac.uk
Frank Pollick: frank.pollick@glasgow.ac.uk

**Participant_Number** – Each participant was assigned a unique number to identify them.

**Initials** – Each participants' initials were anonymised based on their Participant_Number, using letters A-J, with Participant 1 called AA, Participant 2 called BB, ...and Participant 74 called HD.

**Gender** – Participants' self-reported gender.

**Participants_Age** – Participants' self-reported age.

**Nationality** - Participants' self-reported nationality.

**Native_English** – Whether or not participants reported English as their native language.

**Block_Number** – Experimental block in which data was collected. (Range 1-4)

**Interface** – Which interface the participant was randomly assigned during each block. (Options: Control, Iconography, Numerical, Graphical)

**Image_Set** – The set of images randomly assigned during each block. (Options: Image Pool 1, Image Pool 2, Image Pool 3, Image Pool 4)

**File_Number** – Corresponds to the image file taken from the image set which was used for the trial. (Range 1-40)

**Unique_Image_Number** – Assigned to give each image stimuli in the experiment a unique identifier. (Range: 1-160) ***UIN 72 MISSING***

**Trial_Number** – The trial number within each block. (Range: 1-40)

**Cumulative_Trial_Nmbr** – Trial number seen by participant throughout entire experiment. (Range: 1-160)

**Classifier_Correct** – Whether or not the classifier's label correctly described the image contents. (Options: Label Correct, Label Incorrect)

**Image_Clarity** – Whether or not the contents of the image were made difficult to interpret via blurring and occlusion. (Options: Clear, Unclear)

**Trial_Type** – Combination of Classifier_Correct and Image_Clarity. (Options: Correct & Clear, Correct & Unclear, Incorrect & Unclear, Incorrect & Clear)

**Trial_Level** – Numerical version of Trial_Type. (Range: 1-4)

**Classifiers_Label** – Label provided by classifier for image.

**Image_Content** – Contents of image displayed to participant.

**Classifier_Confidence** – Numerical confidence of classifier's top label choice for image.

**Clsfr_Conf_Level** – Confidence level of classifier's top label choice discretised into High (>66%), Low (<33%), or Medium (>33% and <66%).

**Accuracy_Rank** – Rank of each Unique_Image_Number calculated using average Label_Accuracy_Score values given by all participants for image (Range: 1-159)

**Trust_Score** – Measure of how much participants said they trust the classifier (Range: 1-100)

**Label_Accuracy_Score -** Measure of how accurate participants believed the classifier's label was for each image (Range: 1-100)

**Image_Familiarity -** Measure of how recognizable the contents of each image were to participants (Range: 1-100)

**Compliance** – Whether or not participant accepted classifiers label for image (Options: 0, 100)

**Compliance_Outcome** – Text-based version of Compliance (Options: Complies with Classifier, Non-Compliance)

**Trial_Time** – Time spent by participant per trial, measured in seconds.

**Reliability_Block** – Cumulative measure of classifier's reliability for each block of the experiment. Typically starts at 0 or 100 in the first trial, based on whether classifier was correct or not in the trial, and then updates as participants completed more trials in the block. (Range: 0-100)

**TLX_Total** – Participants' total TLX score for each block of the experiment, based on totals for the 6 items asked in the questionnaire. (0-600)

**Mental_Demand** – TLX Item (Range: 0-100) how mentally demanding the task was.

**Physical_Demand** – TLX Item (Range: 0-100) how physically demanding the task was.

**Temporal_Demand** – TLX Item (Range: 0-100) how much time the task demanded.

**Effort** – TLX Item (Range: 0-100) how much effort the task required from participant.

**Frustration** – TLX Item (Range: 0-100) how frustrating the task was to participant.

**Performance** – TLX Item (Range: 0-100) how participants believe they performed in the task.

**Aesthetics** – Asked to participants after TLX how much they liked the design of the interface they worked recently with. (Range: 1-7)

**Propensity_Score** – Participants' total score in the Propensity to Trust Machines Questionnaire. Six questions rated on scale of 1-7. (Range: 6-42)

**Classifier_Helpful** – Debriefing Questionnaire item (Range 1-7) where participants rated whether or not the classifier was helpful in the experiment. <Not at all / A Great Help>

**Classifier_Predictable** – Debriefing Questionnaire item (Range 1-7) where participants rated how predictable or unpredictable the classifier was in the experiment.
<Predictable / Unpredictable>

**Classifier_Specific** – Debriefing Questionnaire item (Range 1-7) where participants rated whether the classifiers labels were too specific, or too general. <Too Specific / Too General>

**Classifier_Characterised** – Debriefing Questionnaire item (Range 1-7) where participants rated whether they thought of the classifier as more of a tool or a teammate.
<Teammate / Tool>

**Classifier_Or_Alone** – Debriefing Questionnaire item (Range 1-7) where participants rated whether they would prefer to work with the classifier again, or if they'd prefer to work alone when classifying images. <With Classifier / Alone>

**Classifier_Or_Human** – Debriefing Questionnaire item (Range 1-7) where participants rated whether they would prefer to work with an automated classifier again, or if they'd prefer to work another human when classifying images. <Computer / Human>

**Preferred_Version** – Participants' explicit preference for one of the 4 classifier interfaces they worked with. (Options: Control, Iconography, Numerical, Graphical).