Consent Form

Consent to Participate in MIT User Study: Heat Maps for Model Understanding

You are asked to participate in a research study conducted by Julius Adebayo, Ilaria Liccardi Ph.D., Hal Abelson and Been Kim, from the Computer Science and Artificial Intelligence Lab at the Massachusetts Institute of Technology (M.I.T.) and Google. You were selected as a possible participant in this study because you are interested in Machine learning. You should read the information below before deciding whether to participate.

PARTICIPATION AND WITHDRAWAL

Your participation in this study is completely voluntary and you are free to choose whether to be in it or not. If you choose to be in this study, you may subsequently withdraw from it at any time without penalty or consequences of any kind. The investigator may withdraw you from this research if circumstances arise which warrant doing so. The investigator will remove you from the study if anomalous behavior is detected in the form of random clicking.

PURPOSE OF THE STUDY

The goal of this prototype study is to assess how well humans can use explanation heat maps as a way to understand machine learning models.

PROCEDURES

If you volunteer to participate in this study, we will ask you to 21 questions relating to the output of a variety of image classification models. Each question should take about 1 min at most to answer (probably less) and this study should not take longer than 25 mins to complete in its entirety.

You will look at a set of images, the output of a machine learning classifier on these images, as well as explanation heat maps. The goal is to assess the machine learning classifier on the basis of the output and the heat maps.

PAYMENT FOR PARTICIPATION

You will receive a $10 Amazon gift card if you successfully complete the study. In addition, we will randomly select 5 participants that complete the survey and award them an additional $30.

CONFIDENTIALITY

To note, we are not collecting any personally identifying information as part of this study. However, any information that is obtained in connection with this study and that can be used to identify you will remain confidential and will be disclosed only with your permission or as required by law. In addition, your information may be reviewed by authorized MIT representatives to ensure compliance with MIT policies and procedures.

Your email will be kept in the server and will be used only to contact you for payment (Amazon Gift card). At the end of your participation this information will be deleted.

IDENTIFICATION OF INVESTIGATORS

If you have any questions or concerns about the research, please feel free to contact Julius Adebayo and Ilaria Liccardi at juliusad@mit.edu & ilaria@csail.mit.edu

EMERGENCY CARE AND COMPENSATION FOR INJURY

If you feel you have suffered an injury, which may include emotional trauma, as a result of participating in this study, please contact the person in charge of the study as soon as possible.

In the event you suffer such an injury, M.I.T. may provide itself, or arrange for the provision of, emergency transport or medical treatment, including emergency treatment and follow-up care, as needed, or reimbursement for such medical services. M.I.T. does not provide any other form of compensation for injury. In any case, neither the offer to provide medical assistance, nor the actual provision of medical services shall be considered an admission of fault or acceptance of liability. Questions regarding this policy may be directed to MIT's Insurance Office, (617) 253-2823. Your insurance carrier may be billed for the cost of emergency transport or medical treatment, if such services are determined not to be directly related to your participation in this study.

RIGHTS OF RESEARCH SUBJECTS

You are not waiving any legal claims, rights or remedies because of your participation in this research study. If you feel you have been treated unfairly, or you have questions regarding your rights as a research subject, you may contact the Chairman of the Committee on the Use of Humans as Experimental Subjects, M.I.T., Room E25-143B, 77 Massachusetts Ave, Cambridge, MA 02139, phone 1-617-253 6787.

☐ **I understand the procedures described above. My questions have been answered to my satisfaction, and I agree to participate in this study. I have been given a copy of this form.**

Demographic Information Block

CONTACT EMAIL *

When entering the email address please be aware that this will be our main point of contact. This email will also be used for issuing payment upon completion of the study.

AGE *

HIGHEST LEVEL OF EDUCATION * (if currently in school please indicate the highest degree received)

- ○ Elementary school only
- ○ Some high school, but did not finish
- ○ Less than high school diploma
- ○ Completed high school
- ○ Some college, but did not finish
- ○ Two-year college degree / A.A / A.S.
- ○ Four-year college degree / B.A. / B.S.
- ○ Undergraduate student
- ○ Some graduate work (e.g. master's student)

- ○ Graduate student
- ○ Professional Degree (e.g. MD, DDS, DVM)
- ○ Master's degree (e.g. MA, MS, MEd)
- ○ Advanced graduate work or Doctorate (e.g. PhD, EdD)
- ○ _____ Other, please specify

GENDER *

- ○ Male
- ○ Female
- ○ Transgender
- ○ Non binary
- ○ Genderqueer
- ○ Intersex
- ○ Questioning
- ○ _____ Please specify
- ○ I rather not answer/disclose

CURRENT EMPLOYMENT STATUS *

- ○ Employed full time (40 or more hours per week)
- ○ Employed part time (up to 39 hours per week)
- ○ Unemployed and currently looking for work

- ○ Unemployed and not currently looking for work
- ○ Student
- ○ Retired
- ○ Homemaker
- ○ Self-employed
- ○ Unable to work
- ○ [ ] Other, please specify

ETHNICITY *

- ○ African American
- ○ Asian
- ○ Hispanic
- ○ Pacific Islander
- ○ White
- ○ [ ] Other, please specify
- ○ I would rather not disclose

Machine Learning Expertise

What is your experience with machine learning (Logistic regression and non-neural networks count!)?

- ○ I have no background in machine learning
- ○ I have limited experience with machine Learning (i.e. I know what machine learning is, but I have not experimented with it).
- ○ I have some knowledge in machine Learning (i.e., I know how to train models)
- ○ I am a machine learning practitioner (I have trained a several models and interacted with them).
- ○ I am machine learning researcher
- ○ [ ] Other

Traditionally, model regularization in machine learning is supposed to help reduce?

- ○ Generalization.
- ○ Overfitting.
- ○ I am do not know the answer

Are you familiar with or have used post-hoc interpretability tools like saliency/attribution heat maps before?

- ○ Yes
- ○ No
- ○ Somewhat

Training Block

Hello and welcome to this user study. In this section we will give you
details of the task you are about to do.

In this study, you will see an artificial agent that classifies pictures of dogs into different breeds. We will call the artificial agent ML through out this task. In addition to the breed of the dog that the ML thinks the picture is, you will also be provided an explanation to see which parts of the image the ML relied on. The image below shows an example ML that takes in a picture of a dog and determines the breed. This is the kind of ML you will see in this study.

Your Role

In this study, you work at a company called KIWI that sells ML to automatically classify dogs into different breeds. Your job title is: "Quality Assurance Tester", and this means that you want to make sure that the ML product that KIWI sells is high quality. Your job requires you to make sure that the ML can properly classify dogs into different breeds.

To help you do your job, you are given a tool that can 'explain' why the ML made its decision. The ML explanation tool highlights the parts of the image that the ML is relying on for classifying the

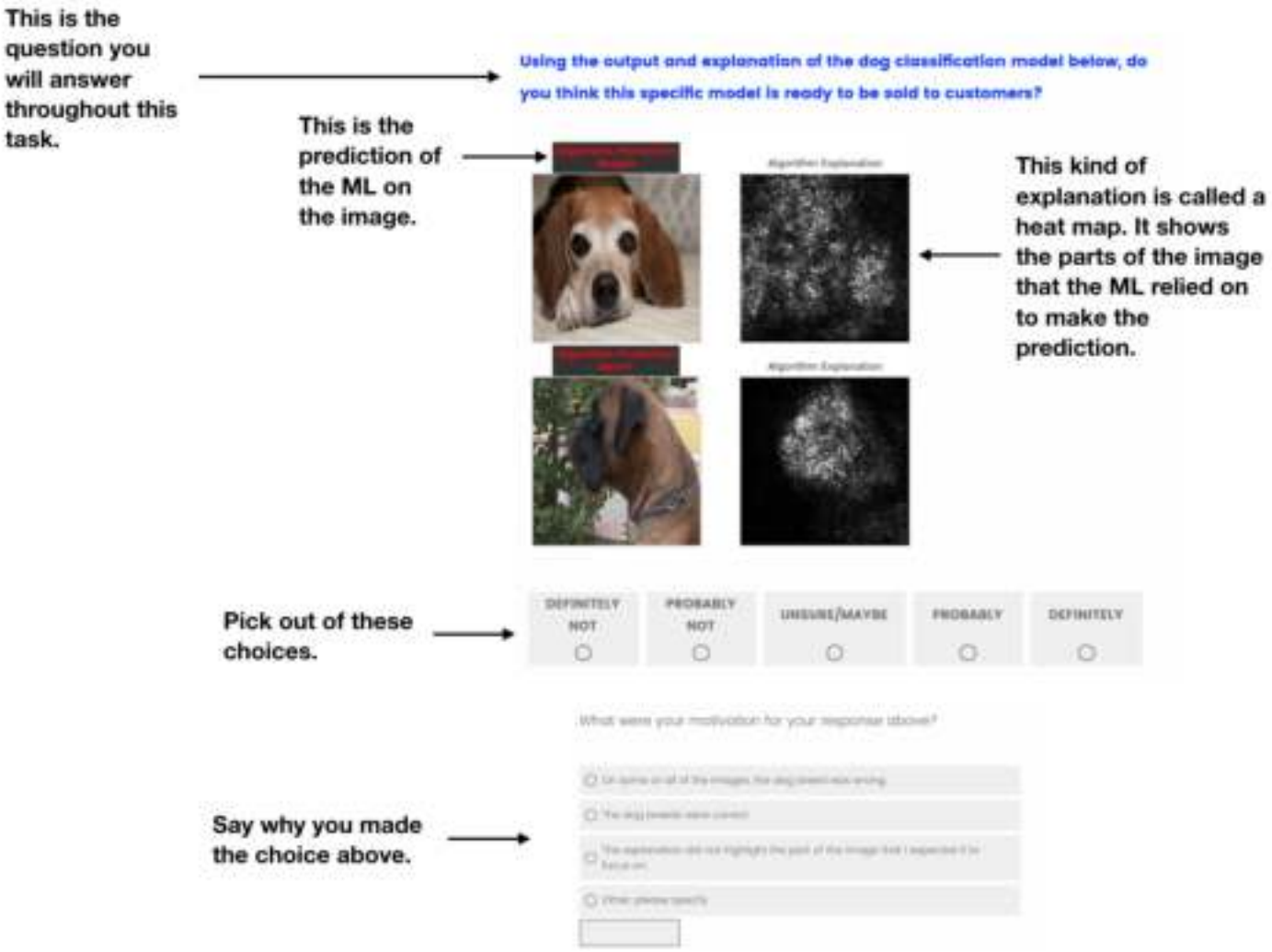
dog image. With this tool, you will decide whether you think the ML is high quality or not. While this tool is not perfect, it may help you make your decision.

Remember - your job is an important step that KIWI relies on to decide whether to sell an ML to external customers.

○ I understand the task and my role.

Overview of the Interface.

The following is an example of the task you will do.

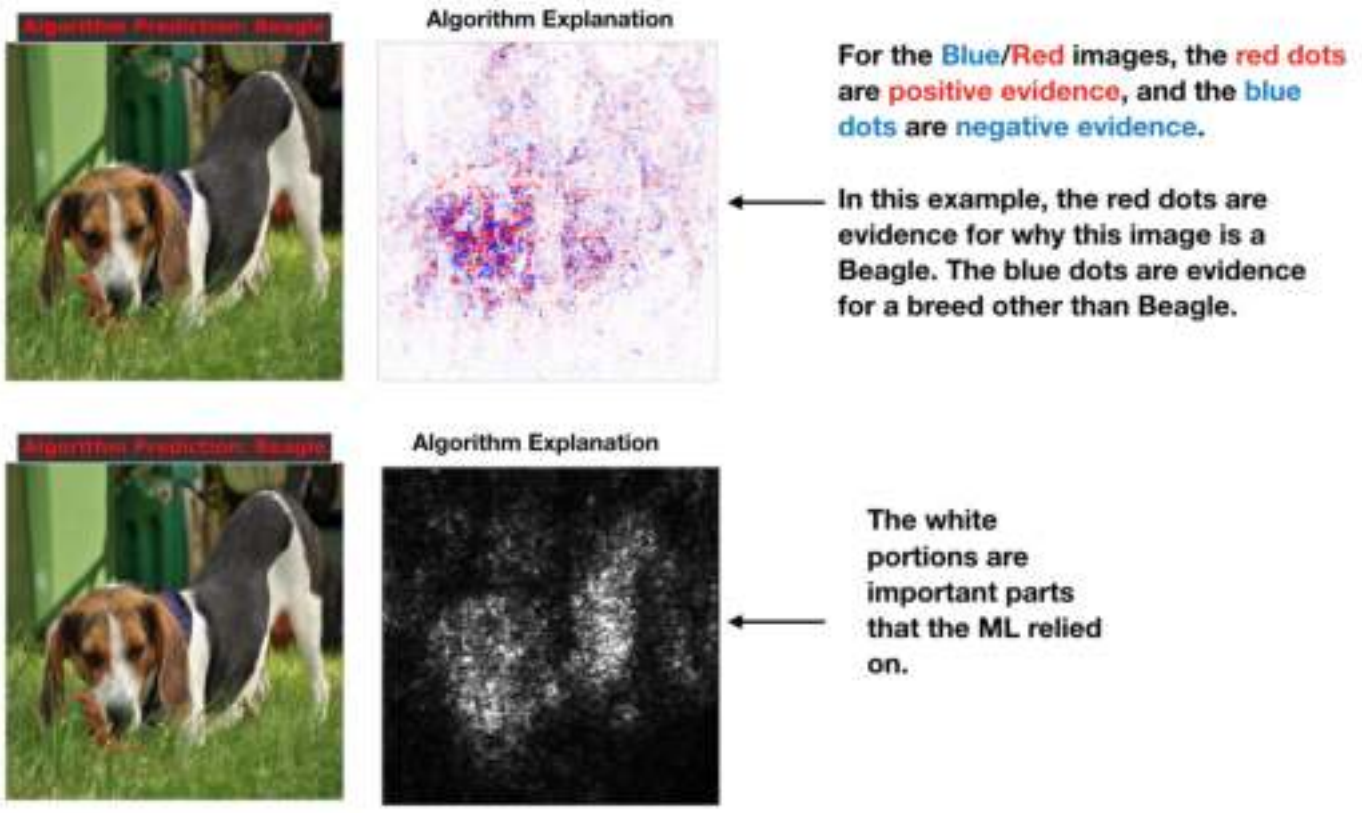


In each task, you will see two dog images. Above each image, you will see, in red text, the ML's prediction for that picture. On the right hand side of each image, you will see the ML explanation for each picture. This explanation highlights the parts of the image the model relies on for its decision. Using the explanations and ML predictions, you will then mark your recommendation on whether this ML is ready for external customers or not.

- [ ] I understand the interface.
- [ ] I understand the explanation tool.
- [ ] I understand what the choices mean.

Overview of the Explanation Types.

You will be seeing two different visualization types. The first explanation type is in black and white. The second explanation type is in red and blue. For the black-white explanation, areas of the image with higher concentration of white are the parts of the image the ML parts more attention to. For the blue and red images, blue areas show negative evidence, while red areas show positive evidence. See the image below for specific examples.



☐ I understand the visualization types.

We will now show you different kinds of dog breeds to familiarize you with dogs.

Beagle.

Boxer.

Chihuahua.

Great Pyrenees.

Newfoundlands.

Pomeranian.

Pugs.

Saint Bernard.

Yorkshire Terrier.

Wheaten Terrier.

Dogs Overview

The study will use image of different dog breeds. What is your familiarity with different breeds of dogs?

In this study, you will be assessing machine learning algorithms that have been built to classify dogs into different species.
You will be shown the outputs of this algorithm along with an explanation tool.
This explanation tool will show a heat map highlighting the parts of the input that the algorithm is relying on.

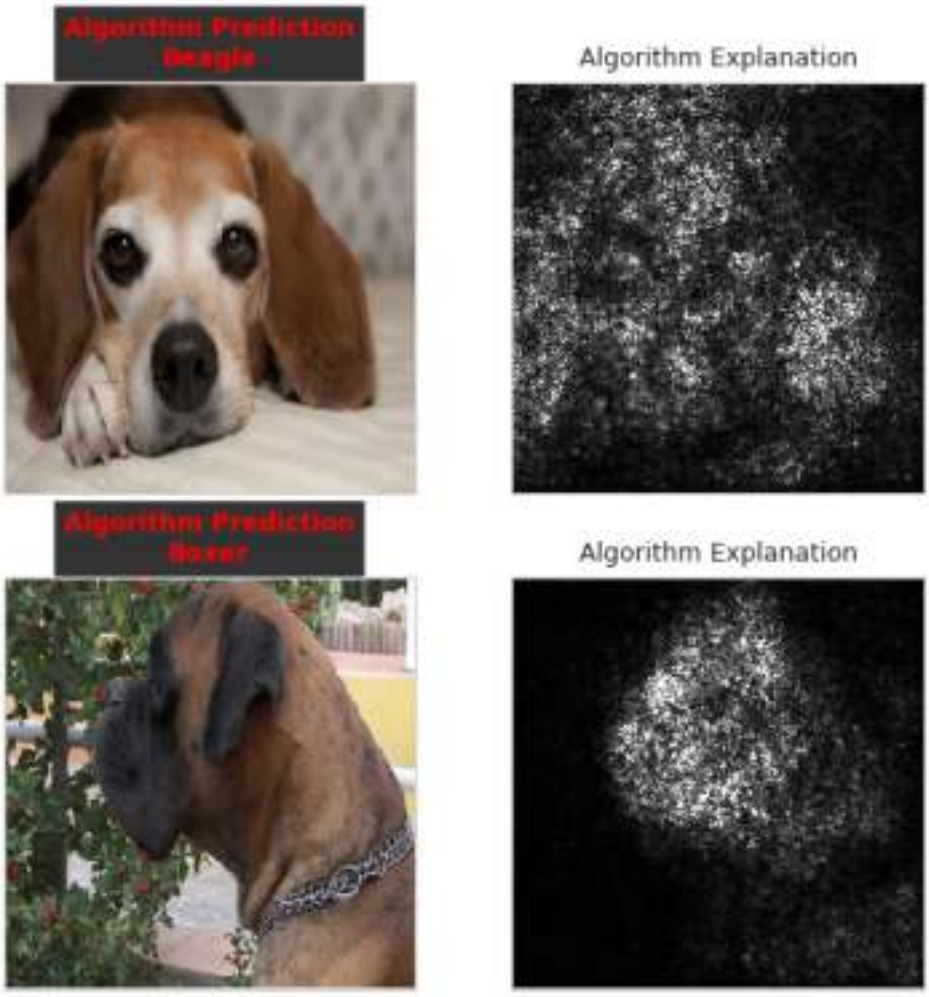
- ○ My knowledge of dog breeds is excellent
- ○ My knowledge of dog breeds is good
- ○ I can recognize few dog breeds
- ○ My knowledge of dog breeds is somewhat limited
- ○ My knowledge of dog breeds is extremely limited.

START

We will now begin the user study, thank you for participating!

G1NS1V1

Using the output and explanation of the dog classification model below, do you think this specific model is ready to be sold to customers?



DEFINITELY NOT	PROBABLY NOT	UNSURE/MAYBE	PROBABLY	DEFINITELY
○	○	○	○	○

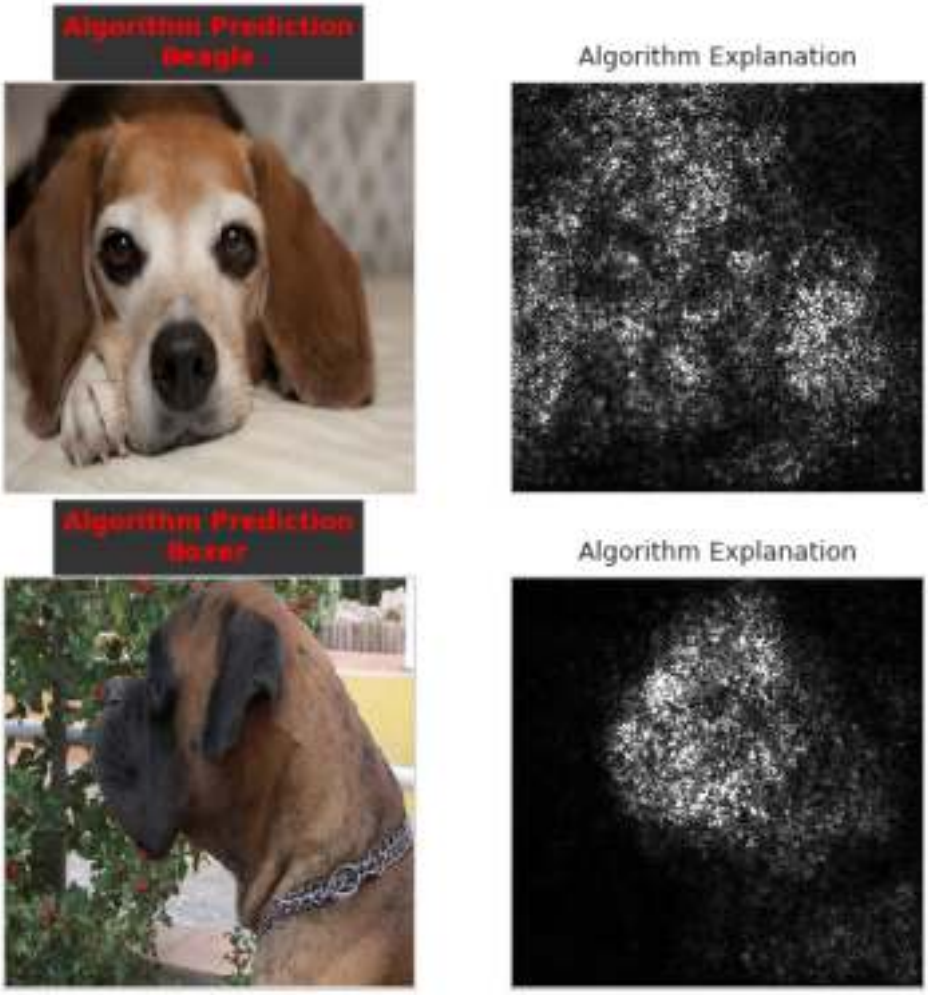
What were your motivation for your response above?

- ○ On some or all of the images, the dog breed was wrong.
- ○ The dog breeds were correct.
- ○ The explanation did not highlight the part of the image that I expected it to focus on.
- ○ The explanation highlighted the parts of the image that I expected it to focus on.

Other, please specify

G1NS1V1_DUP

Using the output and explanation of the dog classification model below, do you think this specific model is ready to be sold to customers?



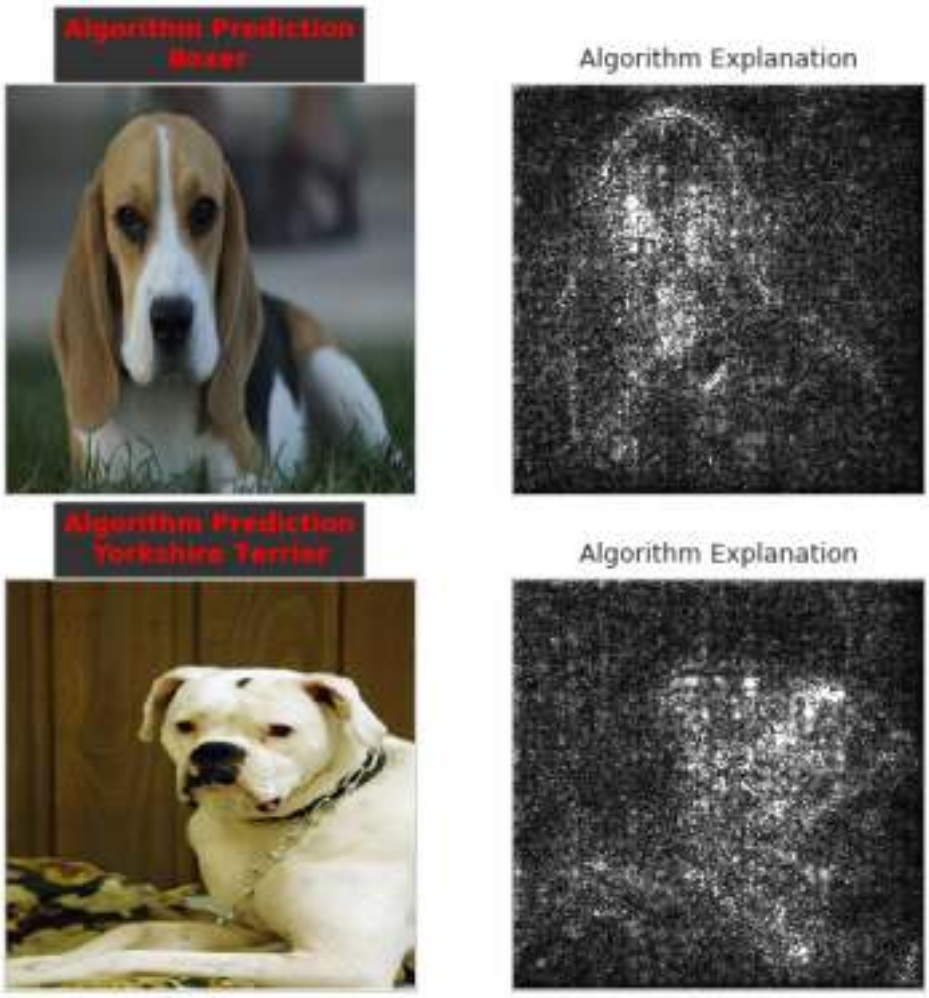
DEFINITELY NOT PROBABLY NOT UNSURE/MAYBE PROBABLY DEFINITELY

What were your motivation for your response above?

- ○ On some or all of the images, the dog breed was wrong.
- ○ The dog breeds were correct.
- ○ The explanation did not highlight the part of the image that I expected it to focus on.
- ○ The explanation highlighted the parts of the image that I expected it to focus on.
- ○ [ ] Other, please specify

G1TLS1V2

Using the output and explanation of the dog classification model below, do you think this specific model is ready to be sold to customers?



DEFINITELY NOT	PROBABLY NOT	UNSURE/MAYBE	PROBABLY	DEFINITELY
○	○	○	○	○

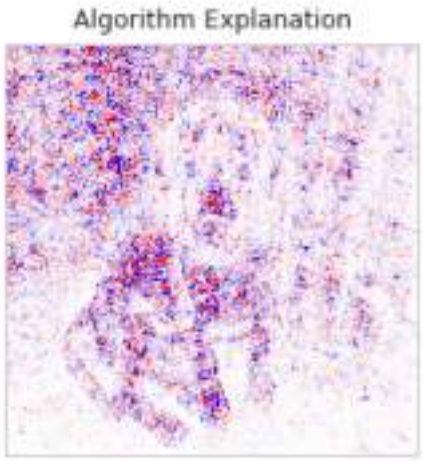
What were your motivation for your response above?

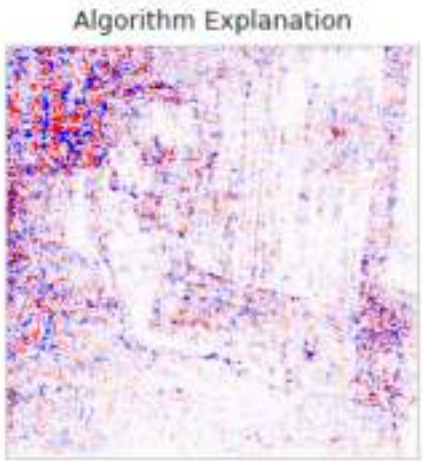
- ○ On some or all of the images, the dog breed was wrong.
- ○ The dog breeds were correct.
- ○ The explanation did not highlight the part of the image that I expected it to focus on.

- ○ The explanation highlighted the parts of the image that I expected it to focus on.
- ○ [ ] Other, please specify

G1HWS1V3

Using the output and explanation of the dog classification model below, do you think this specific model is ready to be sold to customers?





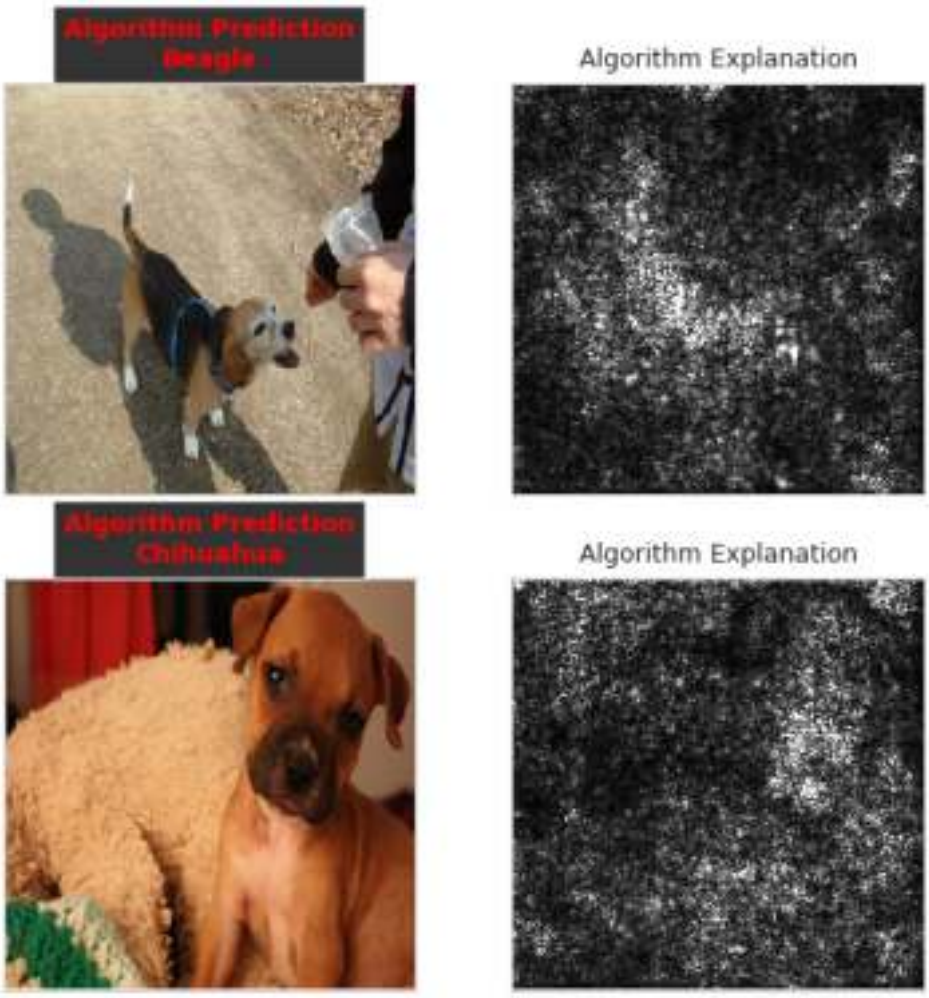
DEFINITELY NOT PROBABLY NOT UNSURE/MAYBE PROBABLY DEFINITELY

What were your motivation for your response above?

- On some or all of the images, the dog breed was wrong.
- The dog breeds were correct.
- The explanation did not highlight the part of the image that I expected it to focus on.
- The explanation highlighted the parts of the image that I expected it to focus on.
- Other, please specify

G1RLS1V1

Using the output and explanation of the dog classification model below, do you think this specific model is ready to be sold to customers?



DEFINITELY NOT	PROBABLY NOT	UNSURE/MAYBE	PROBABLY	DEFINITELY
○	○	○	○	○

What were your motivation for your response above?

- ○ On some or all of the images, the dog breed was wrong.
- ○ The dog breeds were correct.
- ○ The explanation did not highlight the part of the image that I expected it to focus on.