Title

Automatic quantification of disgust taste reactivity in mice using machine learning

Authors/affiliations

Shizuki Inaba, Naofumi Uesaka, Daisuke H. Tanaka

Department of Cognitive Neurobiology, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University (TMDU), 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8519 Japan.

Authors for correspondence:

Naofumi Uesaka; Department of Cognitive Neurobiology, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University (TMDU), 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8519 Japan; Tel +81-3-5803-5445; e-mail: uesaka.cnb@tmd.ac.jp

Daisuke H. Tanaka; Department of Cognitive Neurobiology, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University (TMDU), 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8519 Japan; Tel +81-3-5803-5445; e-mail: dhtanaka.mphm@tmd.ac.jp

Abstract

Disgust represents a quintessential manifestation of negative affect. Prototypical sensory expressions of disgust are triggered by bitter and other unappetizing tastes in human infants, non-human primates, and rodents. Disgust in mice has been quantified through the taste reactivity (TR) test. TR has been video recorded and counted manually to be quantified, requiring a significant amount of time and effort, however. Here we constructed the method to automatically count TR to assess both innate and learned disgust in mice using machine learning. We automatically tracked TR using DeepLabCut as the coordinates of the nose and both front and rear paws. The automated tracking data was split into test and training data, and the latter was combined with manually labeled data on whether or not a TR was present, and if so, which type of the TR was present. Then, a random forest classifier was constructed, and the performance of the classifier was evaluated in the test dataset. Throughout, the total numbers of disgust TRs predicted by the classifier were highly correlated with those counted manually. The present method will facilitate large-scale screening and long-term experiments that require counting numerous TR, which are challenging to conduct manually.

Introduction

Disgust, elicited by the distaste for food in animals, is crucial for avoiding intoxication and infection [1]. In humans, disgust can also be evoked by abstract concepts such as moral transgressions [2]. Abnormal manifestations of disgust towards foods have been observed in patients receiving chemotherapy as a known side effect of various types of anticancer drugs [3-4]. Moreover, previous studies have proposed abnormal disgust in the pathogenesis of various intractable psychiatric disorders, including obsessive-compulsive disorder and eating disorders [5-7]. Therefore, elucidating the neural mechanisms of disgust is essential for reducing drug side effects and effectively managing these disorders.

Disgust is a conscious experience and thus cannot be directly observed. However, taste reactivity (TR), characterized by orofacial and somatic behavioral reactions, has been established as a reliable indicator of disgust [8-10]. TR is widely conserved from human neonates to rodents and distinguishes between disgust and avoidance, whereas neither intake measures nor licking pattern analysis can fully accomplish this differentiation [11-14].

The TR test does have some disadvantages. These include variations in the manual counting of TR in the videos depending on the testers and the significant amount of time and effort required for the counting [13]. These disadvantages may impede the use of the TR test in a wide range of experiments. For instance, research on the parabrachial nucleus (PBN), which is thought to play a key role in certain aversive behaviors, has primarily focused on measuring intake and thus only evaluating avoidance, and not disgust [15]. Utilizing the TR test would provide insight into whether the PBN plays a role in disgust.

We tackled to construct a classifier to automatically count TR to easily assess disgust in mice. To accomplish this, we utilized DeepLabCut [16] as a pose estimation system based on transfer learning with deep neural networks that can track defined body parts of animals without any physical markers and random forest [17] as supervised learning models with an ensemble learning method.

Results

<u>TR test for innate and learned disgust</u>

To assess both innate and learned disgust in C57BL/6 J mice, two groups of mice were prepared and both groups underwent the same surgery to implant the intraoral tube. Group 1 is the Quinine-saline-Quinine group (innate group, n = 13) for innate disgust, and group 2 is the Saccharin-LiCl-Saccharin group (learned group, n = 14) for learned disgust. To allow mice in the learned group to learn disgust, saccharin was injected intraorally, followed by an

intraperitoneal injection of a LiCl solution 2 and 5 days before the TR test, on the day of test, the mice received the saccharin intraoral infusions and the typical orofacial and somatic behaviors of mice during the infusions were video-recorded in a bottom-up view at 30 frames per second (Fig. 1). Mice in the innate group received an intraoral infusion of Quinine, followed by an intraperitoneal injection of saline 5 days before the TR test. In the TR test, the mice received an intraoral infusion of Quinine rather and their behavior was video-recorded.

Manual analysis of TR test

We first manually and quantitatively analyzed video-recorded TRs. We played the recorded videos frame by frame to identify behaviors that could be distinguished as known disgust reactions, such as chin rub, face wash, forelimb flail, gape, and head shake. A total of 97200 frames from all 27 videos were labeled into the five types of disgust reactions (Chin Rub, Face Wash, Forelimb Flail, Gape, Head Shake) and others (Fig. 2A). For five disgust reactions, we also analyzed consecutive frames containing the same reaction as one bout (Fig. 2B). Consistent with a previous study [18], disgust reactions were observed in both innate group and learned group with similar frequencies of occurrence for each disgust reaction (Fig. 2).

Automated body part tracking

Next, to automatically perform quantitative analysis of the video-recorded TRs, we first estimated the mouse posture. We defined five body parts of mice including the nose and the front and rear paws (Fig. 3A). DeepLabCut2.2rc1 [19] was used to develop the tracking framework with the five body parts of mice. A total of 2674 frames from 27 single-trial video recordings were manually annotated. After the model training procedure, pose estimation errors were evaluated. The results showed that the errors were sufficiently small to confirm that the training was complete (Fig. 3B).

Outlier frame detection and correction

We next corrected for mouse posture estimation. Posture estimation is often unreliable due to certain postures (e.g., standing up with its paws against a wall) and rapid movements that are not captured on recording. To prevent incorrect estimation in such cases, we introduced a rule-based correcting procedure that refers to likelihood. Specifically, the following two corrections were made for all frames in videos posture-estimated by DeepLabCut when the estimated likelihood of a certain body part in a certain frame was less than the predetermined threshold value. (1) If the likelihood of the body part in the next frame exceeded the threshold value, the position of the body part in the frame was corrected to the middle of the frame

before and after. (2) If the likelihood of the body part in the next frame was below the threshold, the position of the body part in the frame was corrected to the same position as in the previous frame. In this study, we set the threshold value at 0.6.

Random forest classifiers

Before developing the random forest classifier, the data were split into training (n = 19; innate = 9, learned = 10) and test (n = 8; innate = 4, learned = 4) datasets, randomized equally between the innate mice group and the learned mice group. The classifier was constructed considering the imbalanced presence of frames containing movements of interest and was cross-validated for each mouse in the training dataset to prevent overfitting. After training, we evaluated the performance of the classifier in the test dataset. Throughout, predictions of the total number of individual disgust reactions were compared to manually observed values by human. We found that there was a high correlation between the prediction and the human observation (Pearson's r = 0.98, Fig. 4).

For each behavior, we evaluated the performance of the classifier by calculating positive predictive value (PPV) and sensitivity. PPV was also called precision and expressed by the following equation.

$$\frac{true\ positive}{true\ positive\ +\ false\ positive}$$

Sensitivity was also called recall and expressed by the following equation.

$$\frac{true\ positive}{true\ positive\ +\ false\ negative}$$

In the test set, face wash (sensitivity = 0.69 and PPV = 0.85) (Fig. 5A) and forelimb flail (sensitivity = 0.56 and PPV = 0.74) (Fig. 5B) were predicted with reasonable accuracy, while chin rub (sensitivity = 0.16, and PPV = 0.58) (Fig. 5C), gape (sensitivity = 0) and head shake (sensitivity = 0) were almost undetectable (Table 1). Others (i.e., movements other than the behaviors of interest) were predicted with 0.99 sensitivity and 0.97 PPV.

Discussion

We constructed a classifier using DeepLabCut and random forest to automatically count both innate and learned disgust reactions. The classifier's predicted values showed a strong correlation with human observations. To our knowledge, the present study is the first to automatically assess disgust associated with eating behavior.

Our classifier was constructed in accordance with the published TR test protocols. Therefore, the scores analyzed by our classifier can be compared with the manually counted scores of the previous studies, and we found a very high correlation in the comparison (r =

0.98). Automated analysis methods have been developed in other behavioral tests and have been compared to manual methods. The pup retrieval behaviors in mice were automatically assessed with 0.51-0.90 correlations, using DeepLabCut and machine learning [20]. The number of chemical-induced scratching in mice can be predicted with 0.98 correlations using a convolutional recurrent neural network [21]. Our classifier demonstrated comparable accuracy to these studies.

Our classifier can be built with minimal resources. An inexpensive and commercially available handheld video camera can be used to record mouse behaviors because our classifier required only 30 frames per second for analysis, comparable with previous studies [21-22]. Using videos with lower frame rates not only saves time for pose estimation but also reduces the storage of original videos, ultimately improving the model [23], although some disgust reactions were too fast to be accurately recorded in the video (Figure 5). In addition, the present method required only five tracking points, which is fewer than in previous studies using DeepLabCut [21,24-25]. This reduces the burden of manual annotation and decreases machine training time.

The present study used DeepLabCut for markerless pose estimation since several previous studies have demonstrated its effectiveness in evaluating mouse behaviors as mentioned above. Nonetheless, there are alternative open-source software programs that allow comparable estimations such as DeepPoseKit [26] and SLEAP [27]. Our present classifier could also assess the disgust reactions even if utilizing these programs instead of DeepLabCut. The field of marker-less pose estimation has made remarkable progress in recent years, and more accurate programs will likely become available in the future. Accordingly, automatic behavior evaluation procedures that perform pose estimation in advance, like our present classifier, are expected to improve in accuracy as well since at least some of the failures of our classifier were a result of pose estimation failures (Fig. 5).

Regarding individual disgust reactions, while our classifier predicted forelimb flails and face washes with moderate accuracy, it could not detect gapes and head shakes (Table 1). Despite this limitation, the low prevalence of gapes and head shakes in disgust reactions (Fig. 2) resulted in an overall accuracy comparable to that of human observations (Fig. 4). To improve the detection of specific behaviors, one solution could be to increase the frame rate of the video analyzed and the number of points tracked or to capture video from multiple angles for 3D pose estimation, however, this would come at an added cost.

There is a pioneering study to evaluate responses to stimuli that can cause disgust, using a machine learning-based method for automatically classifying facial expressions in mice [28]. This study found that the facial expression of mice in response to bitter stimuli was distinguishable from their response to other stimuli, resembling the facial expression

observed after the establishment of conditioned taste aversion through exposure to a conditioned stimulus. However, the stimulation protocol in the study makes it difficult to distinguish disgust from other brain functions such as motivation to avoid the stimuli. Thus, it remains unclear what specific brain functions are reflected in the observed facial expressions. Furthermore, the facial expressions detected by this method were challenging to interpret as distinguishable movements relevant to other animal species, including humans. On the other hand, all disgust TR expressed by mice are shared by some animal species, such as rats, monkeys, and human neonates [10]. This fact supports the idea that mouse disgust TR and human disgust emotions share some neural basis. Thus, we believe that TR will continue to be a valuable readout in future research to understand the neural basis of disgust.

Finally, we have created an automated assessment method for disgust TR in mice. The present method allows the evaluation of TR independent of the skill level and bias of the behavior evaluator. In addition, the present method will enable large-scale screening and long-term experiments that require counting numerous TR, which are challenging to conduct manually. Even under different video recording conditions than in our experiments, once a classifier has been created the present method could be adapted. Our proposed method will help the TR test to be conducted in a wide range of experiments. It is expected that the present method accelerates our understanding of the neural basis of disgust and the pathophysiology of various disorders, and advances the development of preventive and therapeutic methods for them.

Materials and methods

Animals

Wild-type C57BL/6J mice were obtained from Japan SLC, Inc. (Shizuoka, Japan). Mice were maintained at 23 ± 1 ˚C under 12 h light/dark cycles (lights on at 8:00 am) and given *ad libitum* access to food and water. All the animal experiments were approved (No. 0150384A, 0160057C2, 0170163C, A2017-194A, A2018-138C4) by the Institutional Animal Care and Use Committee of Tokyo Medical and Dental University and performed in accordance with the relevant guidelines and regulations.

Surgery for intraoral tubing

The surgery for implantation of intraoral tube was performed as described previously [18]. Briefly, a mixture of midazolam (4 mg/kg body weight (BW)), butorphanol (5 mg/kg BW), and medetomidine (0.3 mg/kg BW) was used to anesthetize mice. A curved needle attached to

an intraoral polyethylene tube was inserted from the incision site and advanced subcutaneously posterior to the eye to exit at a point lateral to the first maxillary molar on the right side of the mouth.

TR tests

TR tests were performed as described previously [18]. In brief, on the experimental day (Day) 1, mice in the Sac-LiCl-Sac group received two intraoral infusions of 50 μL of 5.4 mM saccharin solution for 1 min with 1 min interval and then received i.p. injection of LiCl solution (0.3 M, 10 mL/kg BW) just after 2nd saccharin infusion (1st conditioning). The same procedure was repeated on Day 4 (2nd conditioning). On Day 6, the mice received two intraoral infusions of 50 μL of 5.4 mM saccharin solution for 1 minute with a 1 min interval, and behaviors of mice during the infusions were recorded by a digital video camera (HDR-PJ800; Sony, Tokyo, Japan) for the subsequent video analysis. Mice in the Qui-sal-Qui group received intraoral infusions of 3 mM quinine solution in a similar manner to the Sac-LiCl-Sac group on Day 1 and received i.p. injection of physiological saline just after 2nd saccharin infusion. The behavior test for the Qui-sal-Qui group was performed by the infusion of 3 mM quinine solution and video recorded on Day 6.

Manual TR count

Manual video analysis was performed as described previously [18]. To be sure that each component of action contributes equally to the final disgust scores, reactions that occurred in continuous bouts were scored in time bins [10,29]. Components characterized by bouts of longer duration, such as chin rub were scored in five seconds bins (continuous repetitions within five seconds scored as one occurrence) and face wash was scored in two seconds bins. Other three reactions that can occur as a single behavior were scored as separate occurrences (e.g., one forelimb flail equals one occurrence). Individual total disgust reaction count was quantified as the sum of all five reactions during the two minutes of intraoral infusion of test solutions.

Automated body part tracking

The several videos used as the training or test datasets in the current study (n = 13; innate = 6, learned = 7) were from the same videos from an earlier report in which automated tracking was not performed [18]. Initially, 20 frames were selected for each video for training. The provisional coordinates of the mouse body part were estimated according to the temporarily created model, and 40 candidate frames for false predictions were selected for each video and manually annotated again. The same procedure was repeated until the

error was sufficiently small (Fig. 3). Frame choice and training were performed with the default settings of DeepLabCut and trained for 500000 iterations each, achieving a loss of <0.001.

Random forest classification

We extracted the per-frame geometrical 25 features consisting of five distances, ten displacements, and ten displacement changes. In addition, five consecutive frames are considered as the "segment", so each frame had 125 features. Then, we constructed the classifier using caret [30] and Rborist [31] packages in R.

Computer software and hardware

A desktop computer equipped with an Intel Core i7-10700 CPU, an NVIDIA GeForce GTX 1660 SUPER GPU and 16 GB RAM was used for mouse posture estimation in DeepLabCut and data processing in R.

Data availability

The videos, code, and generated datasets used in this study can be provided by the corresponding authors upon reasonable request.

References

[1] Rozin, P. & Fallon, A. E. A perspective on disgust. *Psychol. Rev.* **94**, 23–41 (1987).

[2] Chapman, H. A. & Anderson, A. K. Understanding disgust. *Ann. N. Y. Acad. Sci.* 1251, 62–76 (2012).

[3] Bernstein, I. L. Learned taste aversions in children receiving chemotherapy. *Science* **200**, 1302–1303 (1978).

[4] Jacobsen, P. B. *et al*. Formation of food aversions in cancer patients receiving repeated infusions of chemotherapy. *Behav. Res. Ther.* **31**, 739–748 (1993).

[5] Phillips, M. L., Senior, C., Fahy, T. & David, A. S. Disgust – the forgotten emotion of psychiatry. *Br. J. Psychiatry* **172,** 373–375 (1998).

[6] Knowles, K. A., Jessup, S. C. & Olatunji, B. O. Disgust in anxiety and obsessive-compulsive disorders: Recent findings and Future Directions. *Curr. Psychiatry Rep.* **20,** 68 (2018).

[7] Anderson, L. M., Berg, H., Brown, T. A., Menzel, J. & Reilly, E. E. The role of disgust in eating disorders. *Curr. Psychiatry Rep.* **23,** 4 (2021).

[8] Steiner J. E. The gustofacial response: observation on normal and anencephalic newborn infants. *Symp. Oral Sens. Percept.* **4,** 254–278 (1973).
[9] Grill, H. J. & Norgren, R. The taste reactivity test. I. Mimetic responses to gustatory stimuli in neurologically normal rats. *Brain Res.* **143,** 263–279 (1978).
[10] Berridge, K. Measuring hedonic impact in animals and infants: Microstructure of affective taste reactivity patterns. *Neurosci. Biobehav. Rev.* **24,** 173–198 (2000).
[11] Garcia, J. & Koelling, R. A. Relation of cue to consequence in avoidance learning. *Psychon. Sci.* **4,** 123–124 (1966).
[12] Pelchat, M. L., Grill, H. J., Rozin, P. & Jacobs, J. Quality of acquired responses to tastes by rattus norvegicus depends on type of associated discomfort. *J. Comp. Psychol.* **97,** 140–153 (1983).
[13] Dwyer, D. M. Licking and liking: The assessment of hedonic responses in rodents. *Q. J. Exp. Psychol.* **65,** 371–394 (2012).
[14] Lin, J.-Y., Arthurs, J. & Reilly, S. Conditioned taste aversion, drugs of abuse and palatability. *Neurosci. Biobehav. Rev.* **45,** 28–45 (2014).
[15] Chiang, M. C. *et al.* Parabrachial Complex: A hub for pain and aversion. *J. Neurosci.* **39,** 8225–8230 (2019).
[16] Mathis, A. *et al.* Deeplabcut: Markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21,** 1281–1289 (2018).
[17] Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).
[18] Tanaka, D. H., Li, S., Mukae, S. & Tanabe, T. Genetic access to gustatory disgust-associated neurons in the interstitial nucleus of the posterior limb of the anterior commissure in male mice. *Neuroscience* **413,** 45–63 (2019).
[19] Nath, T. *et al.* Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nat. Protoc.* **14,** 2152–2176 (2019).
[20] Winters, C. *et al.* Automated procedure to assess pup retrieval in laboratory mice. *Sci. Rep.* **12,** 1663 (2022).
[21] Kobayashi, K. *et al.* Automated detection of mouse scratching behaviour using convolutional recurrent neural network. *Sci. Rep.* **11,** 658 (2021).
[22] Jhuang, H. *et al.* Automated home-cage behavioural phenotyping of mice. *Nat. Commun.* **1,** 68 (2010).
[23] von Ziegler, L., Sturman, O. & Bohacek, J. Big behavior: challenges and opportunities in a new era of deep behavior profiling. *Neuropsychopharmacol.* **46**, 33–44 (2021).
[24] Sturman, O. *et al.* Deep learning-based behavioral analysis reaches human accuracy and is capable of outperforming commercial solutions. *Neuropsychopharmacol.* **45**, 1942–1952 (2020).

[25] Nilsson, S. R. O. *et al.* Simple Behavioral Analysis (SimBA) – an open source toolkit for computer classification of complex social behaviors in experimental animals. Preprint at https://www.biorxiv.org/content/10.1101/2020.04.19.049452v2 (2020).
[26] Graving, J. M. *et al.* DeepPoseKit, a software toolkit for fast and robust animal pose estimation using Deep Learning. *eLife* **8,** e47994; 10.7554/eLife.47994 (2019).
[27] Pereira, T. D. *et al.* SLEAP: A deep learning system for multi-animal pose tracking. *Nat. Methods* **19**, 486–495 (2022).
[28] Dolensek, N., Gehrlach, D. A., Klein, A. S. & Gogolla, N. Facial expressions of emotion states and their neuronal correlates in mice. *Science* **368,** 89–94 (2020).
[29] Koizumi, M., Cagniard, B. & Murphy, N. P. Endogenous nociceptin modulates diet preference independent of motivation and reward. *Physiol. Behav.* **97,** 1–13 (2009).
[30] Kuhn, M. caret: Classification and Regression Training. R package version 6.0-93. https://CRAN.R-project.org/package=caret (2022).
[31] Seligman, M. Rborist: Extensible, Parallelizable Implementation of the Random Forest Algorithm. R package version 0.3-2. https://CRAN.R-project.org/package=Rborist (2022).

Figure captions

Figure 1

Experimental setup for TR test in mice: Bottom-up view video recording with oral gavage administration via surgically implanted tube.

Figure 2

The number of each type of disgust reactions.

(A) The number of frames including any disgust reaction manually confirmed.

(B) The number of disgust reaction bouts manually confirmed.

Figure 3

TR test was annotated visible body parts for pose estimation: their errors had decreased and reached plateaus through trainings.

(A) The definition of body parts to estimate mouse posture using DeepLabCut. Only those clearly visible on a bottom-up view were annotated. Scale bar, 1 cm.

(B) RMSE for each body part per training. In this experiment, we recorded the videos with an approximate resolution of 0.13 mm per pixel.

Figure 4

The classifier's predictions were highly correlated with human observations.

The comparison of disgust TR counts in each video between observations and predictions in the test dataset. Pearson's r = 0.98.

Figure 5

Typical examples of positive and false classifications by the present random forest classifier.

The mouse's nose, right front paw, and left front paw are highlighted with green, yellow, and blue crosses, while the right and left rear paws are marked with yellow and blue dots, respectively. Successive images were recorded at 33 milliseconds (ms) (1/30th of a second) intervals. Scale bar, 1 cm.

(A) Examples of the classifier's positive and false classifications of face wash. (Top) A series of frames in which the classifier correctly classified face wash. (Bottom) In this series, the classifier did not detect face wash, and there was a pose estimation error: at t = 0 and 33 ms, the right front paw positions were correctly estimated (red arrow), but at t = 67 ms it was incorrectly estimated (red arrowhead).

(B) Examples of the classifier's positive and false classifications of forelimb flail. (Top) A

series of frames in which the classifier correctly classified forelimb flail. (Bottom) A series of frames in which the classifier did not detect forelimb flail, and there was no pose estimation error.

(C) Examples of the classifier's positive and false classifications of chin rub. (Top) A series of frames in which the classifier correctly classified chin rub. In this series, the left front paw of the mouse was moving so quickly that it cannot be accurately recorded on video: at t = 33 and 67 ms, its positions were correctly estimated (red arrow), but at t = 0 ms its position was incorrectly estimated (red arrowhead). However, the pose estimation was successful. (Middle) In this series, the classifier did not detect chin rub, and there was a pose estimation error: at t = 0 and 67 ms, the right front paw positions were correctly estimated (red arrow), but at t = 33 ms it was incorrectly estimated (red arrowhead). (Bottom) In this series, the classifier did not detect chin rub, and there was no pose estimation error.

Table 1

Observation and prediction about the number of frames including each type of disgust reactions and the other movements.

Figure 1

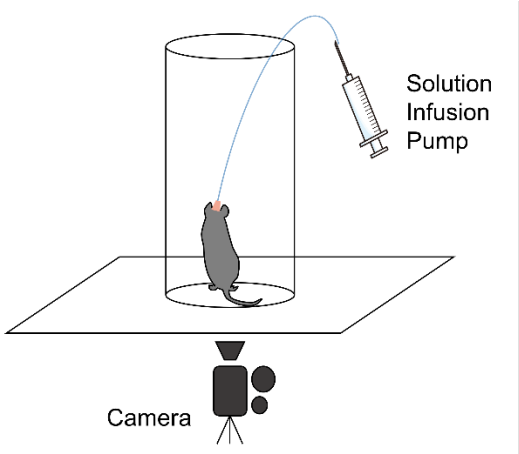


Figure 2

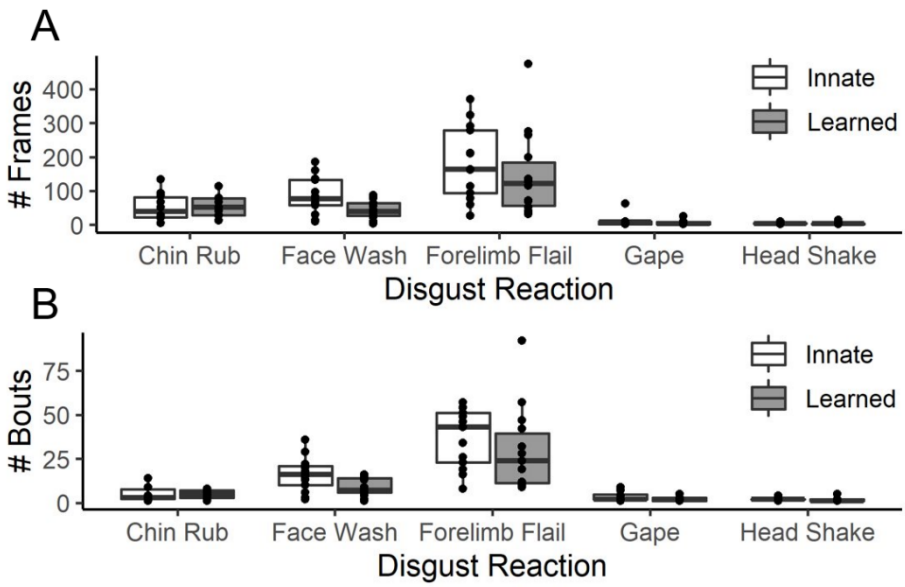


Figure 3

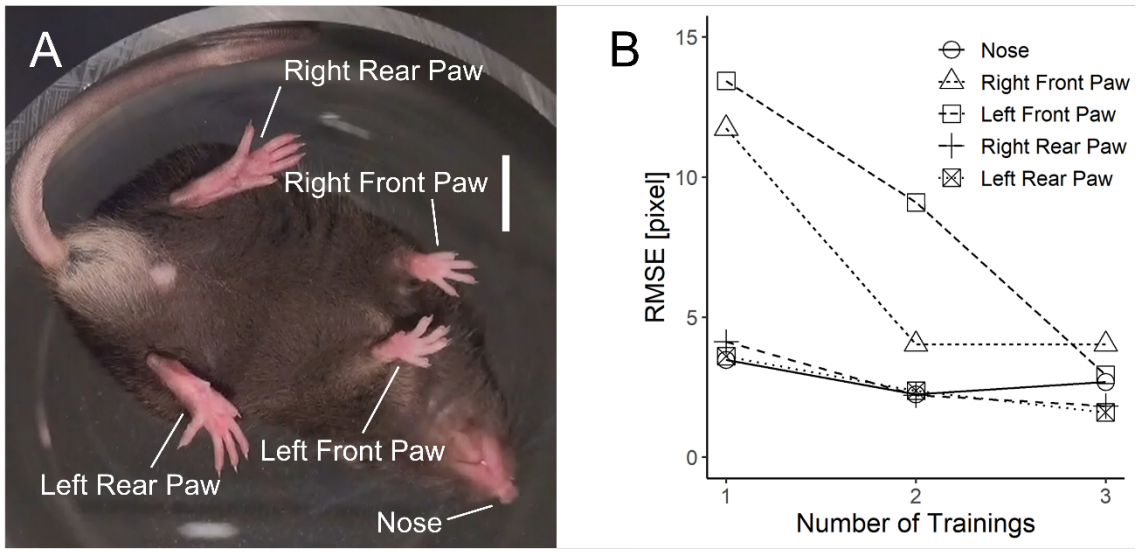


Figure 4

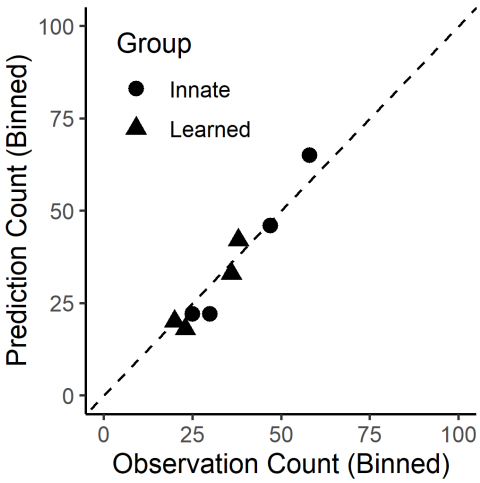
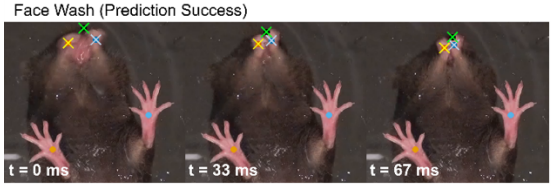


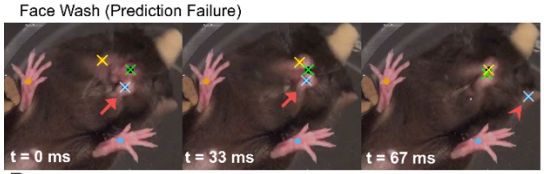
Figure 5

A

Face Wash (Prediction Success)



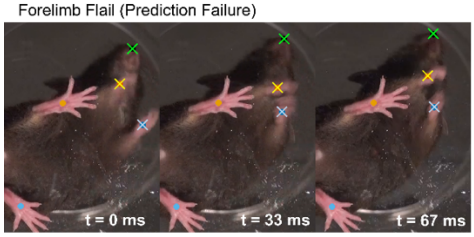
Face Wash (Prediction Failure)



B

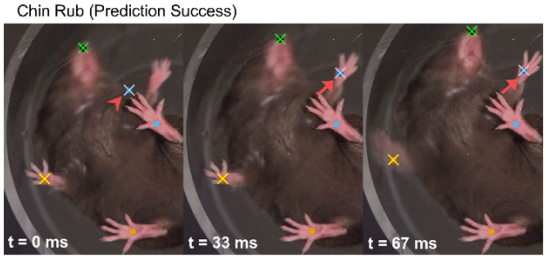
Forelimb Flail (Prediction Success)

Forelimb Flail (Prediction Failure)

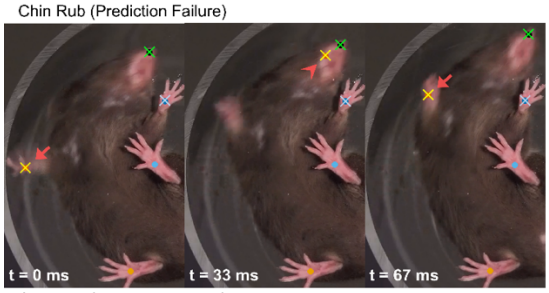


C

Chin Rub (Prediction Success)



Chin Rub (Prediction Failure)



Chin Rub (Prediction Failure)

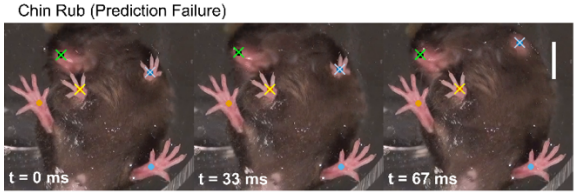


Table 1. Observation and prediction about the number of frames including each type of disgust reaction and the other movements.

	Observation						
Prediction		Chin Rub	Face Wash	Forelimb Flail	Gape	Head Shake	Others
	Chin Rub	45	0	0	0	0	33
	Face Wash	0	283	25	0	0	26
	Forelimb Flail	0	49	496	0	1	121
	Gape	0	0	0	0	0	0
	Head Shake	0	0	0	0	0	0
	Others	236	77	362	22	18	27006