Neuroscience

DeePosit: an AI-based tool for detecting mouse urine and fecal depositions from thermal video clips of behavioral experiments

Reviewed Preprint
v1 • September 2, 2024
Not revised

David Peles ✉, **Shai Netser, Natalie Ray, Shlomo Wagner**

Sagol Department of Neurobiology, Faculty of Natural Sciences, University of Haifa, Israel

Abstract

In many mammals, including rodents, social interaction is accompanied by active urination, also known as micturition, for spatial scent marking. Urine and fecal deposits were shown to contain multiple chemosensory cues carrying information regarding the identity, strain, and social rank, as well as the physiological and hormonal conditions of the individual. Moreover, scent marking was shown to be social context-, state-, and experience-dependent. Thus, analyzing scent-marking activity during social interactions may contribute to understanding the structure of mammalian social interactions in health and disease. So far, however, such analysis faced multiple technical obstacles. Mainly, the commonly used void spot assay relies on detecting urine traces left over a filter paper on which the social interaction occurred; thus, it lacks temporal information and is prone to artifacts such as urine smearing. Recently, several studies employed thermal imaging for spatiotemporal analysis of scent marking, as urine and feces are deposited at body temperature and get rapidly cold afterward. This analysis, however, was done so far manually, which made it time-consuming and prone to bias by the observer. Here, we combine thermal imaging, computer vision tools, and an open-source algorithm incorporating a transformer-based video classifier to automatically detect and classify urine and fecal deposits made by male and female mice during several social behavior tests. We found distinct dynamics for urine and fecal depositions in a test- and sex-dependent manner, suggesting two distinct processes of scent marking in mice. The method and tools presented here allow researchers an easy, efficient, and unbiased spatiotemporal analysis of scent marking during behavioral experiments.

eLife assessment

This manuscript presents a **valuable** machine-learning-based approach to the automated detection of urine and fecal deposits by rodents, key ethological behaviors that have traditionally been very poorly studied. The strength of evidence for their claim, however, that the method provides "easy, efficient, and unbiased spatiotemporal analysis of scent marking during behavioral experiments" is **incomplete**. In particular, there were concerns about the generalizability of the approach, the relatively limited detection capabilities of the method, and a lack of rationale for specific design choices. This manuscript could be of interest to researchers in animal behavior, neuroscience, and automated animal tracking.

https://doi.org/10.7554/eLife.100739.1.sa4

Introduction

In many mammalian species, including rodents, social interactions are accompanied or followed by events of active urination, also known as micturition or voiding activity Arakawa et al. (2008). Multiple studies demonstrated that urine and fecal deposits comprise many chemosensory social cues, which carry information about the individual, such as its species, sex, social rank, and identity, as well as its reproductive and health conditions Bigiani et al. (2005). These chemosensory signals include various metabolites, as well as many proteins such as major urinary proteins Brennan (2004). Thus, by depositing urine spots and feces in its environment, the individual also deposits social information, which may be later perceived by other individuals and modify their future social interactions with this individual Hurst and Beynon (2004). In other words, the use of urine and fecal deposits provides individuals with a way to defend resources such as territory, advertise availability to mates, and communicate with other conspecifics. Specifically, in territorial species, urination is thought to mark the territory of the individual, thus functioning as a spatiosocial scent-marking activity Brennan and Kendrick (2006). Moreover, in rodents, urination activity was shown to be strongly influenced by the individual's internal state, social rank, social context, and previous social experience Desjardins et al. (1973); Hyun et al. (2021). Therefore, assessing scent marking activity may provide valuable information on the individual's social behavior. Specifically, deficits in urine depositing may reflect atypical social behavior in rodent models of various diseases (see Wöhr et al. (2011) for example) and thus may be used for testing potential treatments on such models.

Urination pattern is traditionally analyzed by the void spot assay, which uses filter paper placed on the arena's floor for analyzing urine deposit distribution after the end of the experiment Wolff and Powell (1984); Higuchi and Arakawa (2022). However, this analysis usually lacks the time dimension, is affected by urine smearing and spreading across the arena floor due to the mouse movement (see **Figure 2 d,e**), and is limited in detecting overlapping urine spots. Another limitation is that the filter paper may be torn down by the mouse during the behavioral experiment. Recently, Dalghi et al. (2023) used a setup that includes filter paper on the arena floor, UV light, several cameras, and a manual video annotation to get the time of urination events. Several other studies Verstegen et al. (2020); Miller et al. (2023a) used infrared (IR) cameras for urine detection, as urine is deposited in body temperature and can be seen in the thermal image. However, fecal deposits are also emitted in body temperature, making it difficult to distinguish between feces and small urine deposits using thermal imaging. While using an IR camera solves the issue of temporal analysis, previous studies using this mean relied on manual analysis of the urine spots from thermal video clips, which made the analysis process time-

consuming and subjected to observer bias. To cope with these limitations, we developed an open-source computer vision algorithm to automatically detect urine and feces from thermal video clips. Our detection and classification algorithm is based on a combination of a heuristic algorithm used for the preliminary detection of bright (warm) blobs in the IR video and a trainable video classifier used to classify the preliminary detections as either urine, feces, or background (BG, i.e., not urine or feces). We demonstrate the efficiency of this tool by analyzing the temporal dynamics of both urine and fecal depositions in male and female CD1 (ICR) mice conducting three social behavior tasks. We found that urine and feces depositions show distinct dynamics across the various tests in a sex- and test-dependent manner.

Methods and Materials

Animals

Subject animals were adult (12-14 weeks old) male and female wild-type offspring derived from breeding couples of Gtf2i+/Dup with a CD1 (ICR) genetic background Mervis et al. (2012) mice, bred and grown in the SPF mouse facility of the University of Haifa. Stimulus animals were adult (12-14 weeks old) male and female CD1 mice purchased from Envigo (Rehovot, Israel). All mice were housed in groups of 3-5 in a dark/light 12-hour cycle (lights on at 7 pm), with *ad libitum* food and water under veterinary inspection. Experiments were performed in the dark phase of the dark/light cycle. All experiments were approved by the University of Haifa ethics committee (Reference #: UoH-IL-2301-103-4).

Setup and Video Acquisition

The experimental setup is based on the setup described in Netser et al. (2019). Briefly, a black Plexiglass box arena (37 cm x 22 cm x 35 cm) was placed in a sound-attenuated chamber. A visible light (VIS) camera (both Flea3 and Grasshopper3 models manufactured by Teledyne FLIR were used, both with a wide-angle lens, rate of 30 frames per second, and USB3 interface) and a long wave infrared (IR) camera (Opgal's Thermapp MD with 6.8 mm lens, 384×288 pixels at 8.66 frames per second (FPS)) were placed about 70 cm above the arena's floor. The IR camera was designed to measure human skin temperature and outputs the apparent temperature for each pixel. Raw pixel values were converted to Celsius degrees using the formula supplied by the manufacturer. We acquired the camera videos using custom-made Python software (code is available at: *https://github.com/davidpl2/DeePosit*) that used the manufacturer's SDK (SDK version: EyeR-op-SDK-x86-64-2.15.915.8688-MD). To improve the accuracy of and reduce possible drifts in the measured temperature, a high-emissivity blackbody (Nightingale BTR-03 blackbody by Santa Barbara Infrared, Inc.) was placed in the camera's field of view and was set to 37°C. During analysis, the offset between the blackbody apparent temperature and 37°C was subtracted from the image. To improve image quality, we turned on the camera at least 15 min before the beginning of the experiment (this allows the camera's temperature to get stable). In addition, to reduce pixel non-uniformity, we captured 16 frames of a uniform surface (a piece of cardboard placed in front of the camera) before each test. These images were then averaged, and the average image's mean was subtracted from it to get a non-uniformity image with zero mean. The non-uniformity image was then subtracted from each image in the video to achieve better pixel uniformity.

Social Behaviour Paradigm

We used three distinct social discrimination tests, as previously described in Mohapatra et al. (2024). Briefly, all tests consisted of 15 min of habituation, during which the subject mouse got used to the arena with empty triangular chambers (12 cm isosceles, 35 cm height) located at randomly chosen opposite corners. Each triangular chamber had a metal mesh (18 mm x 6 cm; 1 cm x 1 cm holes) at its bottom, through which subject mice could interact with the stimuli. After habituation, the empty chambers were removed and chambers with stimuli were introduced into

the arena for the 5-minute trial. In the Social Preference (SP) test, a novel (i.e., unfamiliar to the subject mouse) sex-matched stimulus mouse was placed in one chamber, whereas an object stimulus (a Lego toy) was placed in the opposite chamber. In the Sex Prefence (SxP) test, a novel female mouse was placed in one chamber while a novel male was placed in the opposite chamber. In the ESPs test, a novel stressed (restrained in a 50 ml plastic tube for 15 minutes before the test) sex-matched mouse was introduced to one chamber of the arena while a novel naïve mouse was placed in the opposite chamber.

Behavioral Analysis

VIS video clips were analyzed using TrackRodent (*https://github.com/shainetser/TrackRodent*), as previously described in Netser et al. (2017)

Urine and Feces Detection Algorithm

The detection algorithm consists of two main parts. A preliminary heuristic detection algorithm detects warm blobs. These blobs are then fed into a machine learning-based classifier, which classifies them as either urine, feces, or background (i.e., no detection). The algorithm's code is available here: *https://github.com/davidpl2/DeePosit*.

Manual Inputs

A graphical user interface (GUI) was developed in Matlab to support all of the required manual annotations. Each video went through a manual annotation of the arena's floor, the area of the blackbody, and a specification of the first and last frames of both the habituation and trial periods.

These two periods were separated by a 30-second period during which the stimuli were introduced to the arena, which was excluded from the analysis. Also, the arena side of each stimulus (for example, the male and female sides in the SxP test) was defined as the half of the arena close to this stimulus's chamber. To generate the train and test sets, a human annotator manually tagged urine and fecal deposition events in 64 video clips, of which 39 were used for training and 25 for testing. A single click was used to mark the center of each urine or fecal deposit in the first frame where it was visible. The training set included 235 urine annotations and 170 feces annotations. The test set included 56 urine annotations and 90 feces annotations. Additional details can be found in the software's manual.

Preliminary Detection of Hot Blobs

Urine and fecal deposits appear as hot (bright) blobs in the first seconds after deposition. After a cool-down period, which takes about 30-60 seconds for feces and small urine spots and up to ~four minutes for large urine spots, feces and urine appear as dark spots in the thermal image. The preliminary detection relies on these effects (See pseudo-code in Algorithm 1 below). It uses image subtraction to search for hot blobs that appear in the video and cool down later. We generate a background image B_i for each frame F_i to detect new hot blobs. Subtraction of B_i from F_i generates an image in which the mouse pixels and new (warm) urine and feces pixels appear bright. We set B_0 as the per-pixel minimum of the first 20 seconds of video (note that habituation and trial videos are analyzed separately to account for possible minor shifts in the arena's position). We assume that the mouse is brighter than the arena's floor and that the mouse moves during the first 20 seconds, so each pixel will get the arena's floor value at least once during this time.

For $i > 0$ we compute B_i as the minimum of images N_j, $j \in [i - 44, \ldots, i - 36]$ (this roughly matches time range $[i{-}5sec, \ldots, i{-}4sec]$) where N_j is an image in which the mouse pixels were replaced by the last known values from before the time that the mouse occupied these pixels. We set $N_{j<=0} = B_0$.

To compute the mouse mask at frame i, B_{i-1} is subtracted from $F\,i$. The subtraction result is dilated by Matlab's *imdilate* function with a structuring element of a disk of a radius of 2 pixels and then compared against a threshold of 1°C to get a binary mask of the pixels that are warmer than the arena's floor. Connected regions are then computed using Matlab's *bwlabel* function and the connected region with the largest intersection with the arena's floor is considered as the mask of the mouse (denoted M_i).

N_i is then computed by taking F_i values for the pixels outside M_i and taking the values of N_{i-1} for the mouse containing pixels: $N_i = N_{i-1} * M_i + F_i * (1-M_i)$ where * denotes pixel-wise multiplication. The difference image D_i is computed by: $D_i = F_i - max(T, B_i)$ where T is the arena's floor median temperature, computed by $T = median(B_i(AF \,\&\neg M_i \&\neg M_{i-1}))$ where AF is a mask of the arena's floor, & is pixel-wise AND operation and ¬ is pixel-wise NOT operations. Using T prevents higher detection sensitivity in darker regions of the arena floor (regions in the arena's floor that are covered in cooled-down urine appear darker than dry regions of the arena's floor, see **Figure 2e**).

The cooldown rate CD_i is computed by taking the per pixel minimum of the frames in the next 40 seconds following F_i and subtracting it from F_i.

The hot blobs mask BM_i is computed by taking the pixels for which $D_i > 1.1$°C and not included in M_i and M_{i-1} and for which the $CD_i > 1.1$°C and $CD_i > 0.5 * D_i$. We ask for the cooldown to be at least half of the increase in the temperature but not more than that since very large urinations cool down slower and might take more than 40 seconds to cool down fully. We excluded pixels in M_{i-1} and not just M_i since the IR sensor has a response time which might cause pixels included in M_{i-1} to be slightly brighter.

BM_i goes through a morphological close operation using Matlab's *imclose* function with a structure element of a disk with a radius of 4 pixels. This causes any nearby drops of urine to unify to a single detection. Blobs that overlap pixels outside the arena's floor or touch the mouse mask are ignored to avoid detection on darker areas of the mouse (mostly the tail), reflections from the arena's wall, and detections due to a stimulus mouse which sometimes stick his nose throughout the barrier net of the chamber. Also, blobs with a size < 2 pixels or larger than 900 pixels are ignored (pixel size is roughly $0.02cm^2$).

Blobs that intersect previously detected blobs are considered to be the same detection if no more than 30 seconds passed from the last frame in which the previous detection was last detected. A unified detection mask is computed each time a detection is associated with a previous detection. This allows reduction of false alarms which might be caused by the smearing of still-hot urine. If no such intersection exists, a new preliminary detection is added to the list of detections. A blob should be detected in at least two frames to be included in the output detections. The representative coordinate, frame, and mask for each detected blob were chosen by taking the pixel with the maximum intensity inside the blob in all frames it was detected and the mask that matches this frame. Usually, the selected frame for each blob is the first frame of the detection (as the detection cools down, the maximum intensity is usually in the first detected frame). Still, it might be another frame if the detection was partly occluded by the mouse tail or if a second urine event occurred in the same place during the relevant time frame. The output detections are fed into a classifier, which will be described next.

Note that we used relatively low thresholds for the detection (1.1°C) since we wish to detect small urine deposits as well. The detection threshold is slightly higher than the mouse detection threshold (1°C) to avoid false defections on the subject mouse body.

$B_0(p) \leftarrow min_{i\in[1..20FPS]}(F_i(p))$ ▷ Background image at pixel p.
$N_{i<=0} \leftarrow B_0$
Let F_i be the i'th frame in the video
Let AF be the mask of the arena's floor (equals 1 for the arena's floor pixels and 0 elsewhere)
for $i \in [1..n]$ **do**
$M_i \leftarrow BlobWithMaximalFloorIntersection(imdilate((F_i - B_{i-1}) > 1°C, radius = 2))$
$N_i \leftarrow N_{i-1}M_i + F_i(1 - M_i)$ ▷ Mouse pixels are replaced with background pixels
$B_i(p) \leftarrow min_{j\in[i-5sec,...,i-4sec]}N_j(p)$ ▷ Background value for each pixel p
$T \leftarrow median(B_i(AF\&\neg M_i\&\neg M_{i-1}))$ ▷ Median temperature of the arena floor
$D_i \leftarrow F_i - max(T, B_i)$ ▷ Difference image. max operation is pixel-wise
$CD_i(p) \leftarrow F_i(p) - \min_{j\in[i..min(n,i+40*FPS)]} F_j(p)$ ▷ Cooldown in the next 40 sec
$BM_i \leftarrow (D_i > 1.1°CC)\&\neg M_i\&\neg M_{i-1}\&(CD_i > 1.1°C)\&(CD_i > 0.5 * D_i)$ ▷ Hot Blobs Mask
$BM_i \leftarrow imclose(BM_i, radius = 4)$ ▷ Filling small gaps in blobs mask
$BM_i \leftarrow$ blobs in BM_i that are fully inside AF and not adjacent to mouse's mask
with $size \in [2..900]$
blobsList is updated. New blobs are associated with blobs that were detected up to
30 seconds ago if their masks intersect
end for
return blobs in blosbList that were detected in at least 2 frames

Algorithm 1

Preliminary Detection of Hot Blobs

Classifying Preliminary Detections Using an Artificial Neural Network

Preliminary detections are fed to a trained artificial neural network classifier which classifies them as either: *Urine*, *Feces* or *Background* (**Figure 2g**). We relied on the transformer-based architecture proposed by Carion et al. (2020). This architecture was designed for object detection in RGB images. It receives an RGB image as input and outputs a set of bounding boxes around each detected object and the classification of each detection. In brief, this neural network architecture consists of a convolutional neural network (CNN) based on the *ResNet* architecture proposed by He et al. (2016), which serves as the backbone and extracts a set of feature vectors from each location in the input image. The feature vectors are attached with a position encoding, which is a second feature vector that describes the spatial location in the input image, associated with the backbone's feature vector. For each spatial location, the feature vectors from the backbone and the positional encoding are summed and fed into an encoder transformer, which uses an attention mechanism to share information between the feature vectors from various spatial locations. A decoder block is fed with the output of the encoder, and an additional set of vectors is denoted as queries. The decoder uses several layers of self and cross-attention to share information between queries (self-attention) and between the queries and the decoder output (cross-attention). Finally, the encoder outputs a feature vector for each input query. This vector is fed into a feed-forward network (FFN) to compute each query's bounding box and classification. One of the possible classification outputs for each query is "no object". We implemented the popular open-source code published by Carion et al. (2020) with few adjustments. Instead of feeding a single RGB image as input, for each detection in F_i we used a series of 78 grayscale image patches cropped around the detection pixel (65×65 pixels patch) and representing a time window of about [-11sec … 60sec] around the detection. For detection in F_i we used the frames [F_{i-12*8}, F_{i-11*8}, …F_{i-0*8}, …, F_{i+65*8}] for classification. We used this relatively large time window to capture the cooldown of the feces and urine, movement of feces (which are frequently moved by the mouse), or smearing of urine. Additionally, this time window allows for capturing the moment of the deposition of the urine or feces, which sometimes occurs a few seconds before the preliminary detection (since the mouse may fully or partly occlude the detection in the first seconds). Each of the three consequent patches in this set was combined into a single RGB patch and was fed to the backbone. This allows the use of pre-trained backbone weights as well as reduced run-time in comparison to the option of feeding each patch separately to the backbone. Similarly to Carion et al. (2020), each of the backbone's output feature vectors was attached with a positional encoder. However, we adjusted the positional encoding to include additional information on the time of each feature vector (in addition to its spatial location). To do that, we computed time encoding in the same way it was computed by Carion et al. (2020) for encoding the x or y coordinate. To keep the length of the joint position and time encoding the same, we added a fully connected trainable layer that gets the (x,y,t) embedding as input (dim = 128*3=384) and outputs a feature vector with dim=256 which allows using the rest of the neural network and pre-trained weights without additional changes. Lastly, instead of using 100 queries as in Carion et al. (2020), we used just a single query to get just the classification of the input set of patches and disabled the computation of a bounding box. Since our training set is relatively small, we used transfer learning and initialized the learnable weights with the weights published by Carion et al. (2020) (weight file: detr-r50-dc5-f0fb7ef5.pth). We used the dc5 (dilated C5 stage) option proposed by Carion et al. (2020), which increases the spatial resolution of the backbone's output by a factor of 2 as it may be more suitable for classifying small objects, and used ResNet-50 as the backbone.

We trained the classifier using 39 train videos and measured accuracy using an additional 25 test videos. Each video contains a single experiment and includes both the habituation and trial periods (each video lasts roughly 20 minutes). Training database generation included extraction

of: a. Positive examples of urine and feces that were manually marked. b. forty negative examples (labeled as background) per video in randomly selected positions and time (half during habituation and half during trial). c. hard negative examples consist of preliminary detected blobs that are not close in space and time to any manual detection. A preliminary detection in position x_d and time t_d was considered to be close to a manual detection of position x_m in time t_m if $distance(x_d, x_m) < 25pixels$ and $-10sec \le t_d - t_m \le 30sec$. For the positive examples, we augmented the data by a time shift of [-3..6] sec, compensating for possible differences between the manual tagging and the preliminary detection time as well as increasing the training set size. Data augmentation for all examples included a random spatial shift of +-2 pixels, random flip, and rotation of 90, 180, and 270 degrees. Input data was normalized to contain values between [0..255] using linear mapping that mapped 10°C to 0 and 40°C to 255. Values that exceeded 0 or 255 were trimmed. The training was done for 230 epochs with a learning rate of 1e-5 for the backbone and 1e-4 for the rest of the weights and a factor 10 learning rate drop after 200 epochs.

Statistical Analysis

We used a two-sided Wilcoxon rank sum test (Matlab's *ranksum* function) for all pairwise comparisons. Rank sum p-value equal to or smaller than 0.1, 0.05, 0.01, 0.001 was marked with #, *, **, ***, respectively. In addition, since some of the data is zero-inflated (many mice do not deposit urine or feces in the relevant measured period), we used a two-way chi-square test to compare the distribution of zeros and non-zeros in the male group vs. the female group in **Figure 5** and in **Figure 5— figure Supplement 1**. The two-way chi-square test was implemented using Matlab (see code in Listing 1). P-value equal or smaller than 0.1, 0.05, 0.01, 0.001 was marked with !, +, ++, +++, respectively, and was mentioned to the left side of the ranksum p-value symbol (i.e, the notation +/** means that two-way chi-square test resulted in p-value<=0.05 and the rank sum test resulted in p-value <= 0.01). For the habituation vs. trial comparison (**Figure 4a-b** and **Figure 4—figure Supplement 1**), and the side preference analysis (**Figure 6**), mice with zero urine detections across all periods of the same test were ignored. The same was done for the feces analysis. Lastly, we used Matlab's *kruskalwallis* function for the Krusukal-Wallis test, which was used to examine the effect of test type (SP, SxP, ESPs) on the dynamics of the urine and feces (***Table 1*** and **Supplementary Table 1**). Additional statistical data for the figures is available at *https://github.com/davidpl2/DeePosit/FigStat*.

Results

Social Discrimination

We analyzed the time subject mice spent investigating each stimulus during the various tests (**Figure 1**), using the video clips recorded via the visible light (VIS) camera. Both male and female subject mice showed the behavior expected from CD1 mice, as previously described by us Kopachev et al. (2022). For males, we found a significantly higher investigation time towards the social stimulus as compared to the object in the SP test, towards the opposite sex as compared to the same sex stimulus mouse in the SxP test, and towards the stressed as compared to the naïve mouse in the ESPs test. Females showed similar behavior, except for the SxP test, where they showed no preference for any of the two stimuli. In accordance with our previous study Netser et al. (2017), in all cases, the preference towards a given stimulus was reflected only by long (> 6s), but not by short (≤ 6s) investigation bouts (**Figure 1**). Thus, in terms of social behavior, the subject mice behaved as expected.

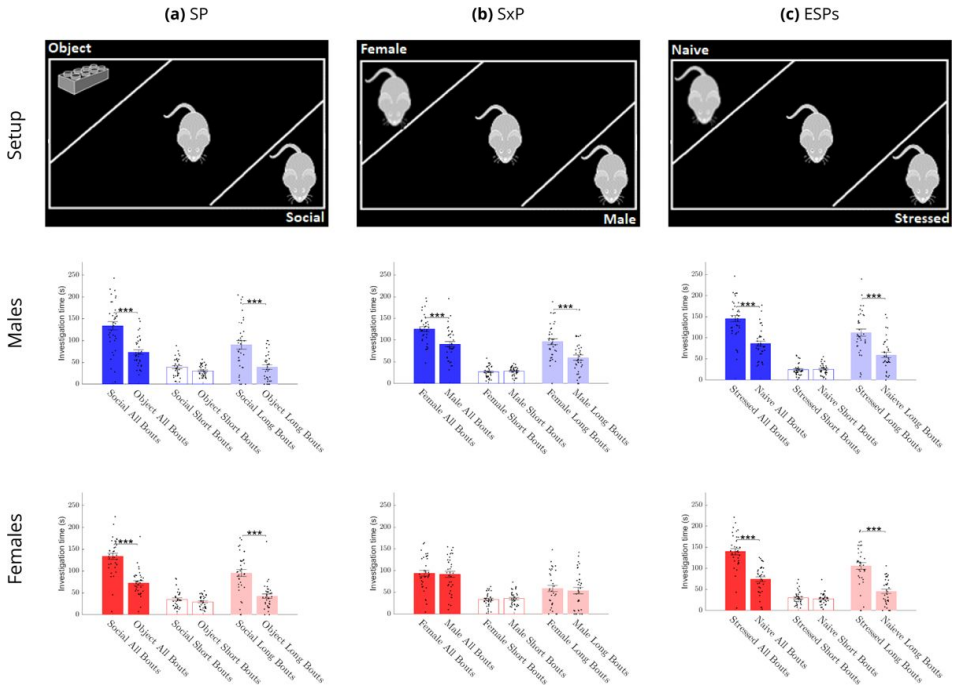


Figure 1.

Investigation time and bout duration across sexes and tests. The first row shows the arena's setup, while the second and third rows show the mean (±SEM) time dedicated by male (n=36, blue bars) and female (n=35, red bars) mice to investigate each stimulus during the (a) SP, (b) SxP and (c) ESPs tests. The two leftmost bars in each panel show the total investigation time, while The two middle bars show the time spent on short (≤6s) investigation bouts, and the two rightmost bars in each panel show the time spent on long (>6s) investigation bouts.

Urine and Feces Detection Accuracy

The experimental setup, including the VIS and IR cameras, is schematically shown in **Figure 2a**. Unlike the VIS camera (**Figure 2b**), the IR camera captured the warm urine and feces drops soon after they were deposited (**Figure 2c**). This allowed us to overcome several caveats of the void spot assay. For example, we could tolerate smeared urine spots (**Figure 2d-e**) and identify the time of each deposition event (See Video 1, Video 2 and **Figure 2**—video 1). For the generation of training and testing data sets, a human annotator manually tagged urine and fecal deposition events in 64 video clips, of which 39 were used for training the model and 25 for testing. The detection algorithm (termed DeePosit) consists of two main parts. A preliminary heuristic detection algorithm detects warm blobs (**Figure 2f**). These blobs are then fed into a machine learning-based classifier (**Figure 2g**), which classifies them as either urine, feces, or background (i.e., no detection). For more details, see the Methods section. The confusion matrix for the testing dataset showed 78.6% recall rate for the detection of urine deposits and 83.3% recall rate for fecal deposits **Figure 2h** and **Figure 2—figure Supplement 1**. A manually tagged urine or fecal deposition was considered correctly detected if an automatic detection with the same label exists in a distance of up to 20 pixels and in a time difference of up to 15 seconds. The spatial tolerance is required due to ambiguity in the tagging process of urine, as some manual taggers might mark large spots or long traces of urination differently, mainly if a trace of urine includes several spots (see **Figure 2d** for an example). Moreover, the detection algorithm might unify adjacent urine spots while being tagged as two different urine depositions by human annotators (see **Figure 2**—video 1 and Supplementary Figure 1 for examples). The temporal tolerance is required since the mouse body may cover the deposit or be very close to it for a while, thus delaying the time the preliminary detection algorithm detects it. Note that for large urine deposits, the classification accuracy is higher (**Figure 2—figure Supplement 1**), probably because it is more distinguishable from fecal deposits, which are always small. See **Figure 2**—video 1 and Supplementary Figure 1b for examples of mistakes made by the detection algorithm in the test videos, which are further discussed in the Discussion section.

Distinct Dynamics of Urine and Fecal Depositing Activities

Figure 3a,b shows the raw results of urine and fecal deposit detection by the DeePosit algorithm for each mouse as a function of time across all three tests, for both male (Blue symbols) and female (red symbols) subject mice. The symbols representing the various types of deposits are also labeled with black dots, according to the arena side of each deposition (see figure legend). These raw results were further analyzed by computing the average number of urine or fecal deposition events, as well their average area (cm^2), per minute **Figure 3c,d**, which was calculated since urine deposit size might vary significantly between distinct events and conditions Wegner et al. (2018). Generally, the event rate and deposition area showed similar trends. As apparent, urine and feces depositing activities showed distinct dynamics: feces depositing activity showed a single clear peak in all cases in the early stage of habituation. In contrast, urine deposition was characterized by two peaks, which were not visible in the SP test but appeared in the SxP and got even stronger in the ESPs test. The first urination peak occurred in males at the early habituation stage, parallel to the peak in feces deposition. The second urination peak occurred in both males and females after stimuli insertion into the arena. For statistical analysis of these dynamics, we compared the mean urine and fecal deposition rate between three periods: the beginning of habituation (habituation minutes 1-5), the end of habituation (habituation minutes 11-14), and the trial - after stimuli introduction (trial minutes 1-4) (**Figure 4a,b**). For both males and females, we found a significantly higher level of fecal deposition at the beginning of habituation than at the end in all tests (except for SP in females, when only a trend was observed). In contrast, a similar comparison of urine deposition showed that its level was significantly higher during early habituation than at the end only for males in the SxP and ESPs tests. A similar elevation in urine deposition, specifically during the SxP and ESPs tests, was observed during the trial, compared to

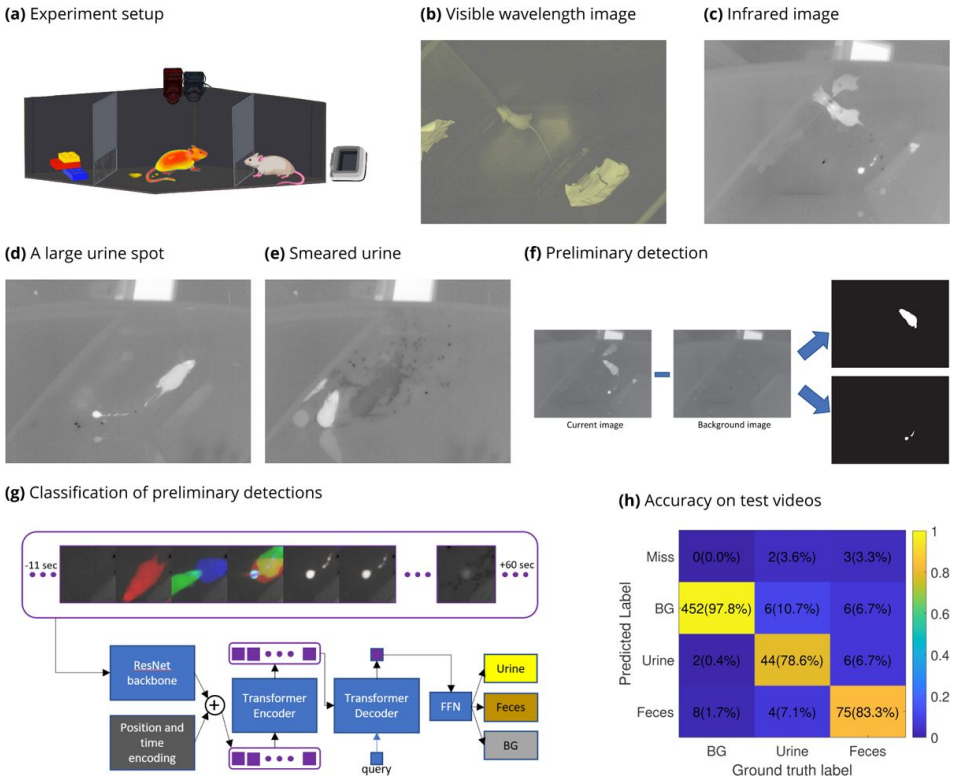


Figure 2.

The experimental setup and analysis method The experimental setup (a) includes a visible light (VIS) camera, an infrared (IR) camera, and a blackbody set to 37°C. VIS (b) and IR (c) images that were captured at the same moment, a short time after a urine deposition, exemplify that, as the urine is still warm, it appears as highly contrasted blob in the IR image but not in the VIS one. Large urine spots, such as the one shown in (d), may be smeared across the arena's floor (e), which is one limitation of the use of filter paper for quantifying urination at the end of the experiment. The preliminary detection algorithm is based on subtracting a background image from each frame in the video (f), which allows the detection of hot blobs reflecting the animal itself and urine and feces deposits. The detected blobs are then classified using a transformer-based artificial neural network (g), which gets as its input a time series of patches cropped around the detection and provides its classification as an output. Each three patches in that time series are merged into a single RGB image (see methods). In the confusion matrix presenting the accuracy of the full pipeline for test videos (h), the "Miss" row counts the events that were not detected by the preliminary hot blobs detection and, hence, were not fed to the classifier. The BG (background) column counts the number of automatic detections for which no matching manually tagged event exists in the relevant space and time window. See Methods for more details **Figure 2—figure supplement 1**. Accuracy for small and large detections. **Figure 2**—video 1. Video for the events in the confusion matrix. Each part of the video matches a cell in the confusion matrix (h) and shows the events included in this cell (up to 48 events). Each event is shown in a 65×65 pixel window from −11 seconds before the event to +60 seconds afterward (similar to the classifier input). The video shows both the manual annotation and the automatic detection that was matched with it (shown side by side). Note that there are no automatic detections for the "Miss" row of the confusion matrix and no manual annotation for the BG column of the confusion matrix. The video plays at X3 speed.

the habituation end, for both males and females. Interestingly, we found an opposite trend for fecal deposits, with a significant decrease in fecal deposition rate during the trial in the ESPs test for males and the SxP test for females (**Figure 4a,b**). Similar results were found for urine and fecal deposit areas (**Figure 4—figure Supplement 1**). Moreover, similar trends were observed when the proportion of mice actively depositing urine or feces during each period was calculated for each case (**Figure 4c**). These data reveal distinct dynamics for urine and feces deposition activities during the various tests in a sex- and test-specific manner.

Sex-Dependent Differences During the Habituation Period

We used two types of statistical tests to compare males and females. A two-sided Wilcoxon rank sum test (significance marked by *) was used for all pairwise comparisons. In addition, since some of the data is zero-inflated (many mice did not deposit urine or feces at all during the relevant period), we used a two-way chi-square test (significance marked by +) to compare the distribution of zeros and non-zeros in the male group vs. the female group. A test-dependent significant difference between males and females was found in the first 5 minutes of habituation (**Figure 5a**). On the first day of experiments (the SP test), males and females showed a low urination rate at the first 5 minutes of habituation, with no significant difference between them. However, in the next two testing days (SxP and ESPs tests), when the mice were already familiar with the arena (as they had already gone through the SP test) we found a significantly higher rate and area of urine deposition in males compared to females (**Figure 5a** and **Figure 5—figure Supplement 1 a**). This difference is more significant in the last experiment (the ESPs test), where we also found a significant difference between males and females in the distribution of urinating mice (mice with at least one urine detection in habituation minutes 1-5). As for fecal deposition, males showed a trend towards a higher level in this period across all tests. During the last stage of habituation, we found a significant difference between males and females only for the ESPs test, with males showing higher levels of both urine and fecal deposition rate (**Figure 5b**), as well as area (**Figure 5—figure Supplement 1 b**).

Sex-Dependent Differences During the Trial Period

For statistical comparison between males and females during the trial, where an initial peak was observed in some cases (**Figure 3c-d**), we divided it into two periods, the first minute and minutes 2-4, and averaged the results of each period separately. As apparent in **Figure 5c-d** and **Figure 5**— figure Supplement 1 c-d, the urine deposition results of the trial's first minute were similar to those of the early stage of habituation, with no difference in the SP test, which was conducted first, and a significantly higher level of urination events for males vs. females in the SxP and ESPs tests, which were conducted later. For trial minutes 2-4, we found a significant difference between males and females only for the ESPs test. No difference was observed for fecal deposition in any of the tests trial periods.

Male Urine and Fecal Deposition Rates are Test-Dependent

Since the data so far suggest a dynamic change from the SP to the SxP and ESPs tests specifically for males, we checked the effect of test type (SP, SxP, ESPs) on the urine and fecal deposition dynamics using Kruskal-Wallis test ***Table 1*** and **supplementary Table 1**. Males' urine and fecal deposition rates (***Table 1***) and area (***Supp Table 1***) showed a significant effect of the test type, with urination showing this effect during early habituation and trial, while fecal deposition showing such effect at both early and late habituation, but not during the trial. No significant effect was found for females.

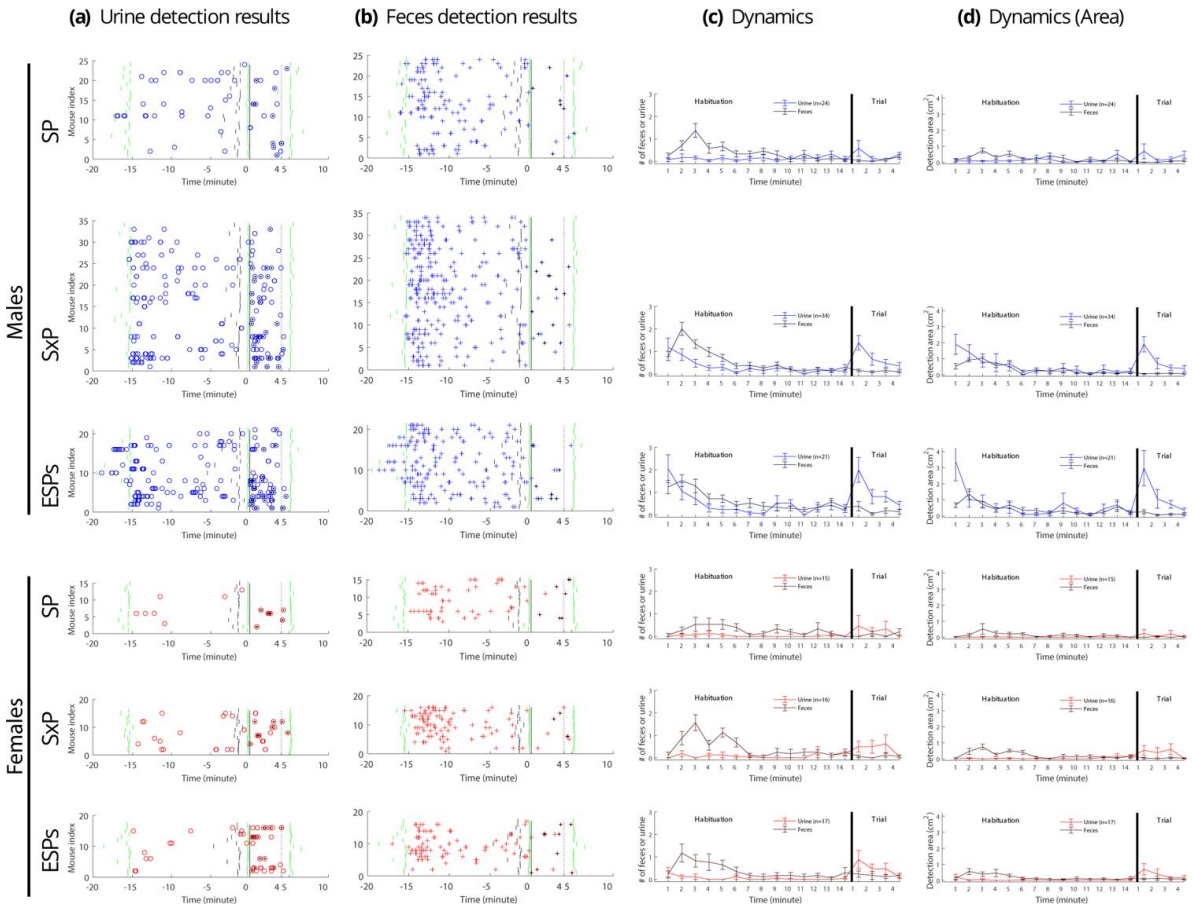


Figure 3.

Urine and fecal deposition detection results across tests.

Each circle represents a single detection of urine deposition (**a**), while each + represents a single detection of fecal deposition (**b**). Green lines mark the start and end of habituation and the end of the trial. The vertical black line at time=0 marks the stimuli's introduction and the trial period's start. The vertical dotted line marks 4 minutes after the beginning of the trial. The short vertical black lines mark the end of minute 14 of the habituation. A black dot in the center of a circle or a + sign marks that this detection is on the side of stimulus1 (preferred stimulus), defined as the social stimulus in the SP trial, the female in the SxP trial, and the stressed mouse in the ESPs trial. Dynamics graphs show mean rate (**c**) and mean area (**d**) per minute of urine and feces. Error bars represent standard error.

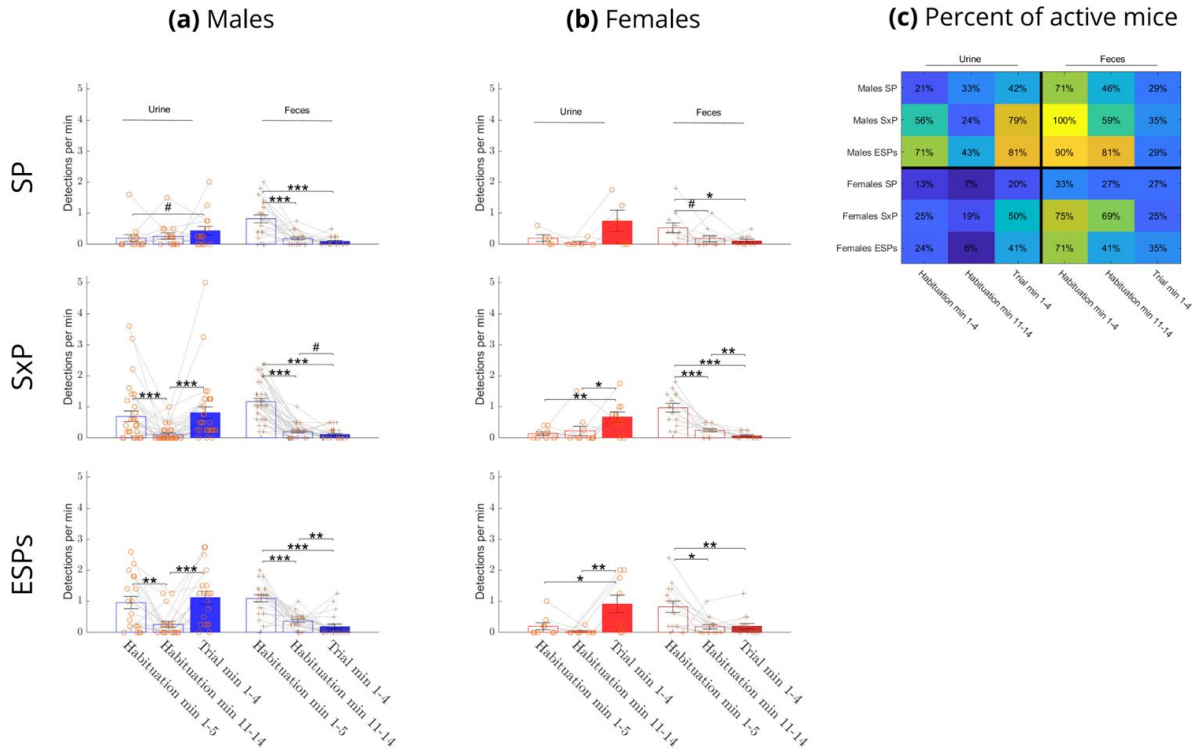


Figure 4.

Comparison between test periods.

The mean rate of urine and fecal deposition during habituation start (minutes 1-5), habituation end (minutes 11-14), and trial (minutes 1-4) for males (**a**) and females (**b**). (**c**): Percent of active mice (mice with at least one detection) across tests during the same periods as above. **Figure 4—figure supplement 1**. Urine and fecal depositions area during habituation start, habituation end, and trial.

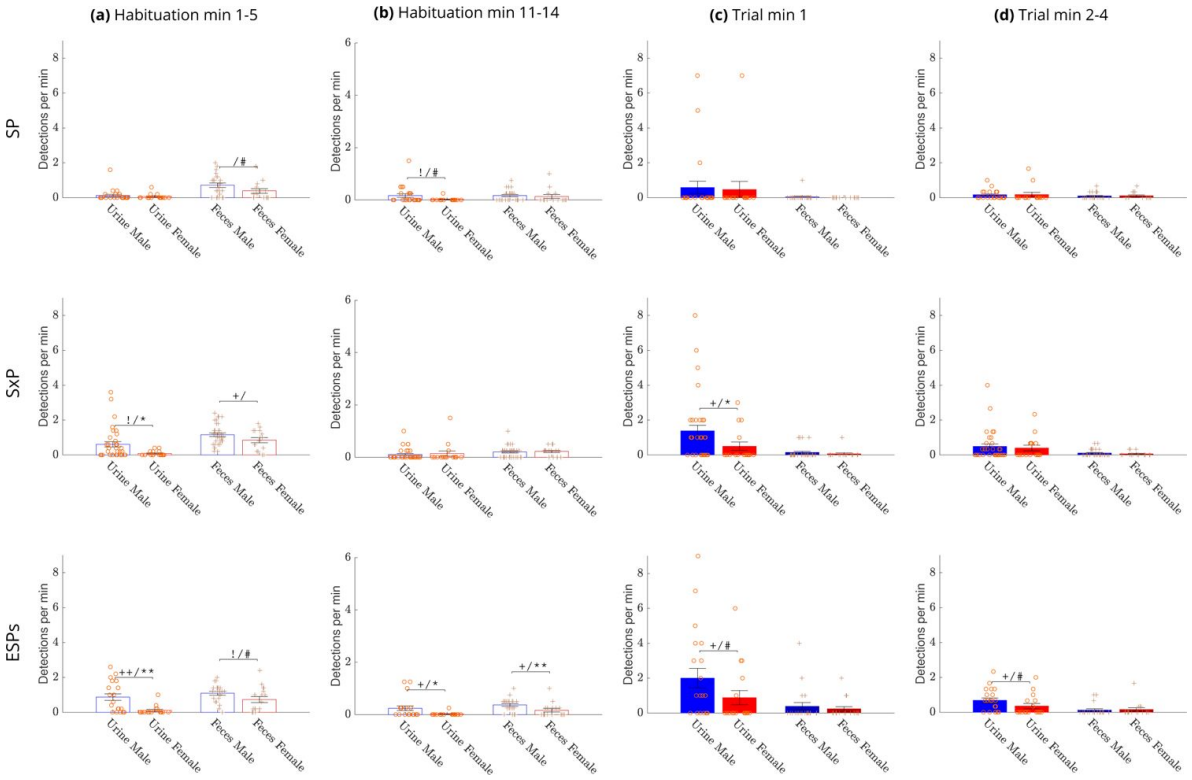


Figure 5.

Comparison of deposition rates between sexes.

The mean rate of urine and fecal depositions in males (blue bars) vs. females (red bars) during early (minutes 1-5) and late (minutes 11-14) minutes of habituation and during the first minute and minutes 2-4 of the trial. A significant difference between the mean rate of urine or fecal depositions (Wilcoxon rank sum test) is marked with * (or # for 0.05<p-value ≤0.1), and a significant difference in the distribution of non-depositing animals (Chi-square test) is marked with + (or ! for 0.05<p-value ≤0.1). **Figure 5 —figure supplement 1. Comparison of deposition areas between sexes.**

Figure 6.

Urine and fecal deposition side preference.

A comparison of the mean ±SEM rate ((**a**) and (**b**)) and area ((**c**) and (**d**)) of urine (two left bars in each panel) and fecal (two right bars in each panel) depositions made by male (blue bars) and female (red bars) subject mice in each side of the arena, for all three tests. Rank sum p-value equal to or smaller than 0.1, 0.05, 0.01, 0.001 was marked with #, *, **, ***, respectively

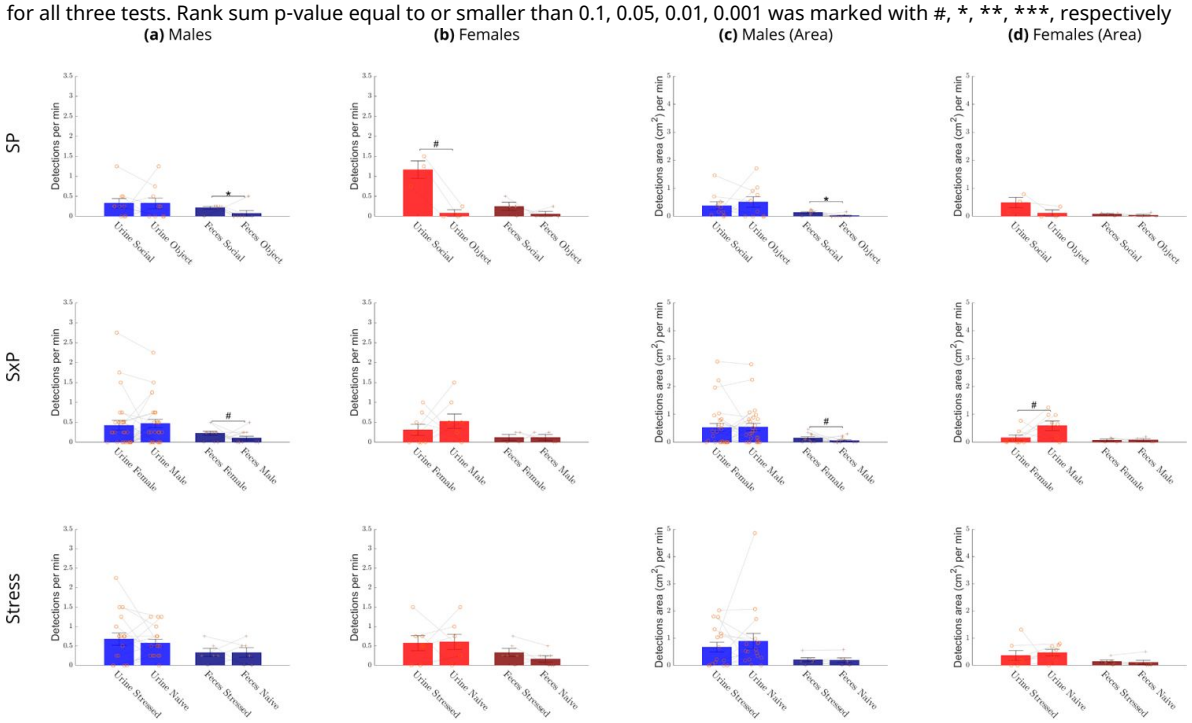


Table 1.

The effect of the test (SP, SxP, and ESPs) on the urine and fecal deposition rates. Kruskal-Wallis test was used to check if the test type affects the rate of urine or fecal depositions.

Measurement	Habituation1-5	Habituation11-14	Trial1	Trial1-4
Male #Urine	0.0012**	0.3497	0.0031**	0.0100**
Female #Urine	0.7931	0.3729	0.3161	0.3568
Male #Feces	0.0232*	0.0079**	0.2711	0.7462
Female #Feces	0.0830#	0.1272	0.1853	0.8354

Males Make More Fecal Depositions at the Social Side During the SP Test

Finally, we found that males make more fecal depositions at the arena side of the social stimulus during the trial period of the SP test (p=0.029 for # of detections and p=0.017 for detections area), a tendency that became trendy in the SxP test and disappeared in the ESPs test (**Figure 6a,c**). For females, we found only a few trends but no significant difference (**Figure 6b,d**). Interestingly, while males spent more time on the social side during the SP trial (**Figure 1**), they did not deposit more urine on that side of the arena.

Discussion and Limitations

Here, we present a new algorithm and an open-code trainable computational tool for detecting and classifying urine and fecal deposition events from IR video clips. This algorithm allows detailed characterization of small rodents' urine and fecal deposition dynamics during social behavior experiments. The advantage of this tool is that it is automated, thus creating rapid and unbiased analysis of urination and fecal deposition events and areas with a high temporal and spatial resolution. Specifically, combining our algorithm with an IR camera for thermal imaging of behavioral experiments, as we conducted here, can replace the void spot test, which usually lacks any temporal resolution and is prone to mistakes caused by urine smearing and filter-paper tearing. Finally, our algorithm allows analysis of fecal deposition behavior, which was rather unexplored so far, but may contribute to scent marking behavior, as discussed below. Our algorithm uses thermal video clips generated by an IR camera placed above the arena and does not require a thermal camera placed below a clear arena floor, as used by several recent papers (see Keller et al. (2018) for example). Thus, it can be employed with standard experimental setups, such as those used for the three-chamber test. We believe the computational tool and experimental method presented here can be useful for a detailed characterization of social behavior in mice, including the context of mice models of autism spectrum disorder and other social behavior-related health conditions. It may also help investigate urination and fecal deposition activities in other scientific contexts unrelated to social behavior. Our experimental setup is cheap and easy to assemble, and the detection algorithm can run on a standard PC with a GPU card.

Analysis of the mistakes made by the algorithm in the test set (see **Figure 2**—video 1) raised several limitations, which might be addressed in future work. Urine or fecal depositions must be fully visible and not partially occluded by the mouse when the deposit is still warm. Partial occlusion or a close adjacency between the mouse and the urine or fecal blob might cause the mouse mask to overlap the mask of the urine or fecal deposit, thus preventing their detection. All of the "miss" events in the test videos (two urine and three fecal depositions) were close to the mouse for a long period after their depositing. A wrong classification of urine as fecal deposition occurred four times in the test set. In all these events, the urination spot was small (and therefore harder to distinguish from feces). In two of these events, there was a second urine or fecal deposition in a nearby location after the first urine deposition. We hypothesize that such an event may cause the classifier "to think" that a shift in the location of the deposit (which is common in the case of feces) has occurred. Classification of background as feces occurred eight times in the test set. In seven of these events, the mistake was due to feces that were shifted by the mouse to a new location while they were still warm. Classification of background as urine occurred two times in the test set. One of these events was caused by a few drops of urine splashed by the stimulus mouse. In the second case, true urination appeared a few seconds after the preliminary detection in a nearby location, which was visible in the input image patches fed to the classifier. Future work might improve accuracy by extending the training set and including more challenging examples. Another future work may use a trainable detection and segmentation algorithm instead of heuristic preliminary detection. Note that our classifier currently does not get as input the mask

of the preliminary detection, making the classification task harder when there are adjacent urine and feces events. An end-to-end trainable detection, segmentation, and classification pipeline might address these limitations but might require a larger training set.

We validated our method and algorithm using experimental results from social discrimination tests conducted by male and female CD1 mice. We showed that in this context, there are distinct statistically significant dynamics of urine and fecal deposition activities across the habituation and trial stages, and these dynamics are sex- and test-dependent. Both males and females showed higher levels of fecal depositing at the early stage of the habituation phase **Figure 4a-b**. This tendency may reflect the higher level of anxiety expected at the beginning of the habituation phase. Still, it may also serve as a scent-marking activity that labels the arena as a familiar place for the subject animal. The latter explanation is supported by the fact that the peak in fecal deposition activity was not reduced from the first-day test (SP) to the third-day test (ESPs) when the subject is expected to be less anxious due to the familiar spatial context. In contrast to fecal depositing, urine deposition activity at the beginning of the habitation phase was test-dependent. While no peak was observed during the SP test, the first time the animals were exposed to the experimental arena, it was observed in the second test (SxP) and got even stronger in the last test (ESPs). This development was statistically significant in males but not in females. Since these changes occur during the habituation phase, before the introduction of the stimuli to the arena, they cannot reflect the type of test and thus seem to be induced by the order of the experiments. Notably, similar dynamics across experimental days were previously reported using the void spot assay for a different strain of mice Keil et al. (2016). This suggests that the induction of urination activity at the early stage of the habituation phase in males represents territorial scent-marking activity, which is positively correlated to the higher familiarity experienced by the subject in the arena as the experiments progressed between days. It should be noted the early peak of urination upon entering an environment was reported by a recent study using a thermal camera for manual analysis of urination activity Miller et al. (2023b). A second peak of urination activity was observed at the beginning of the trial period, after stimuli insertion to the arena. This was observed in both males and females, but the test type significantly affected it only in males. In this case, we cannot dissect the effect of test type from the test order, as the urination activity occurred after stimuli insertion and, hence, may be induced by the presence of specific social stimuli. Since the subjects are already habituated to the arena at this stage, the elevated urination activity seems to serve as part of the subjects' social behavior, most probably as a territorial scent-marking behavior. Interestingly, we did not observe a consistent spatial distribution of the urine or fecal deposits between the arena sides of the preferred and non-preferred stimuli. This seems to contradict a recent study Miller et al. (2023b), that reported opposite bias towards familiar vs. unfamiliar stimuli in losers vs. winners wild-derived mice following a social contest. This contradiction may be due to the distinct mouse strains, distinct contexts of social behavior, or different times with stimuli used by both studies.

Overall, the novel algorithm and software presented by us here enable a cost-effective, rapid, and unbiased analysis of urine and fecal deposition activities of behaving mice from thermal video clips. The algorithm is trainable and may be adapted to various behavioral and experimental contexts. Thus, it may pave the way for integrating this important behavioral aspect in analyzing rodents' social and non-social behaviors in health and disease.

Acknowledgements

We want to thank Yaniv Goldstein, Janet Tabakova, and Shorook Amara for their help annotating the videos and Sara Sheikh for drawing the experiment setup illustration. This study was supported by ISF-NSFC joint research program (grant No. 3459/20), the Israel Science Foundation (grants No. 1361/17 and 2220/22), the Ministry of Science, Technology and Space of Israel (Grant No. 3-12068), the Ministry of Health of Israel (grant #3-18380 for EPINEURODEVO), the German

Research Foundation (DFG) (GR 3619/16-1 and SH 752/2-1), the Congressionally Directed Medical Research Programs (CDMRP) (grant No. AR210005) and the United States-Israel Binational Science Foundation (grant No. 2019186).

Appendix 1

Appendix 1—table 1.

The effect of the test on the urine and feces area. Kruskal-Wallis test was used to check if the test type (SP, SxP, and ESPs) affects the area of urine or feces.

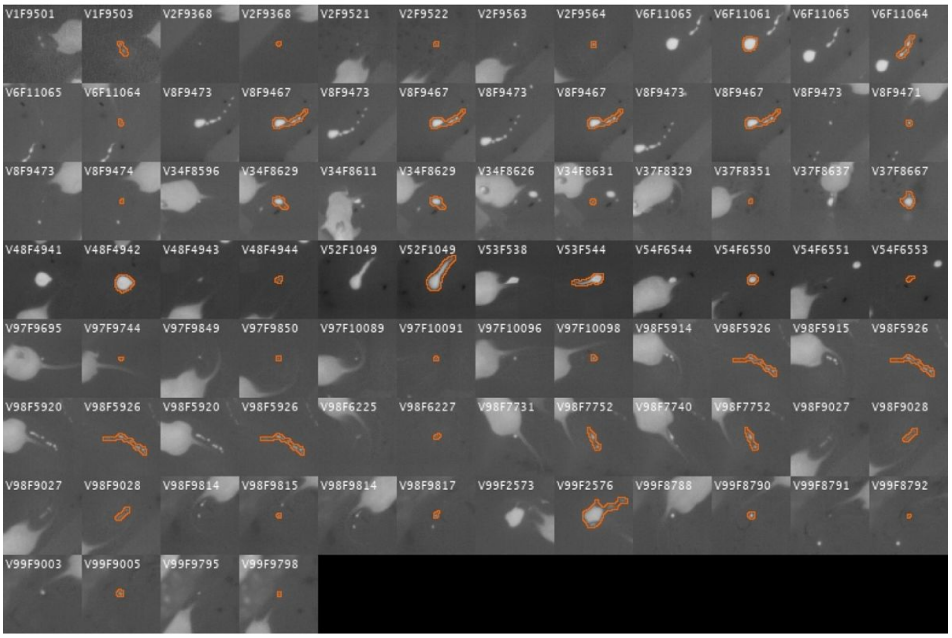
Measurement	Habituation1-5	Habituation11-14	Trial1	Trial1-4
Male Urine Area	0.0005***	0.3297	0.0018**	0.1047
Female Urine Area	0.7822	0.3524	0.3044	0.2595
Male Feces Area	0.0091**	0.0671#	0.2350	0.6452
Female Feces Area	0.2198	0.2333	0.2090	0.9427

Listing 1.

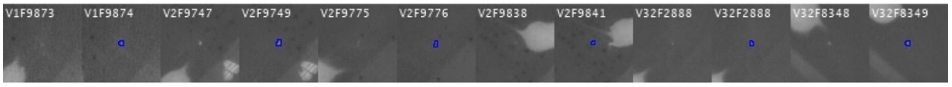
Code for computing Two Way Chi-Square Test which was used to compare the distribution of active mice (with at least one detection) in males vs females.

```
%Compute two way chi square test for 2x2 table
%
%Hypothseis H0: there is no relation between gender and the distribution of zeros.
%Hypothseis H1: there is a relation between gender and the distribution of zeros.
%
%Inputs:
% valsMales   - a vector of length 2 that contains: [zeros count, non zeros count] for
    males.
% valsFemales - a vector of length 2 that contains: [zeros count, non zeros count] for
    females.
%
%Outputs:
% pVal - p value. A value lower than 0.05 suggests that the hypothesis H0 should be
    rejected.
% chiStat - statistic of the chi square test.
% df - degree of freedom (equals 1 for 2x2 tables).
%
function [pVal,chiStat,df] = TwoWayChiSqrTest(valsMales,valsFemales)

if length(valsMales)~=2 || length(valsFemales)~=2
    error('input vectors should have length=2')
end
sumMales   = sum(valsMales);
sumFemales = sum(valsFemales);
sumAll     = sumMales+sumFemales;

sum1 = valsMales(1)+valsFemales(1);
sum2 = valsMales(2)+valsFemales(2);
expectedFreqMales   = [sumMales  *(sum1/sumAll), sumMales  *(sum2/sumAll)];
expectedFreqFemales = [sumFemales*(sum1/sumAll), sumFemales*(sum2/sumAll)];

chiStatMales   = sum((valsMales  -expectedFreqMales  ).^2 ./ expectedFreqMales);
chiStatFemales = sum((valsFemales-expectedFreqFemales).^2 ./ expectedFreqFemales);

chiStat = chiStatMales+chiStatFemales;
df = 1;
pVal = 1-chi2cdf(chiStat,df);
```

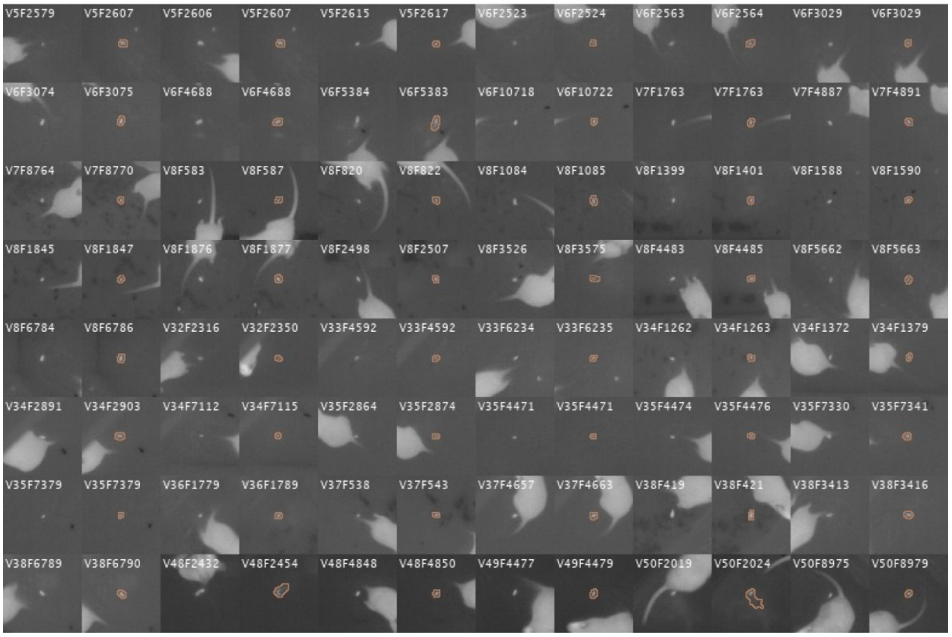
(a) Correctly detected urine



(b) Urine that was classified as background



(c) Correctly detected feces



Appendix 1—figure 1.

Examples of detections in test videos. (a,b,c) are screenshots taken from ***Figure 2*** —video 1. (a): Examples of urination events that were detected and classified correctly. Each pair of columns includes a ground truth detection (to the left) next to the matched automatic detection (to the right), which includes the mask of the detected blob. The overlaid text mentions the video index and the frame index. (b): Urination events that were wrongly classified as background. Note that all of these urine spots are very small. (c): Fecal depositions that were detected and classified correctly.

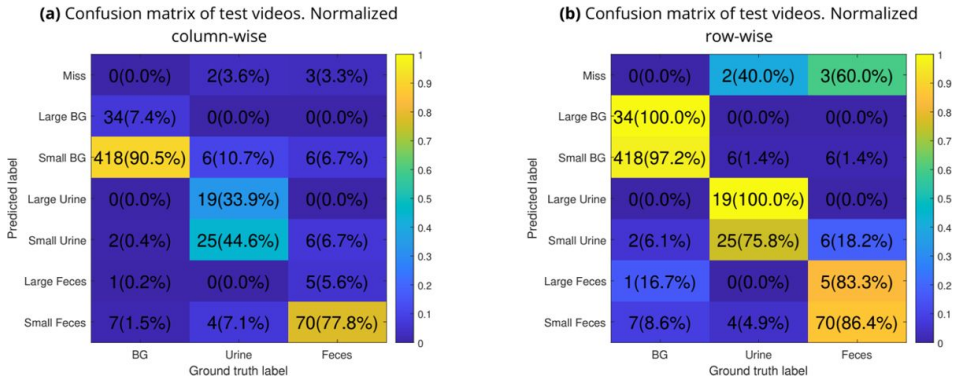


Figure 2—figure supplement 1.

Accuracy for small and large detections. (a,b) Confusion matrices on test videos with separation between large and small automatic detections. The threshold for large detections is an area of $1cm^2$ which is 47.3 pixels. Shown percents sum to 1 for each column in (a) and each row in (b). The *Large Urination* class is correct in 100% of the cases in which it was reported by the classifier while *Small Urination* is correct in only 75.8% as shown in (b). Most of the confusion between feces and urine spots is for small detections: 7.1% of the Ground Truth (GT) urine events were classified as *Small Feces* while 0% as *Large Feces* as shown in (a). Also, 6.7% of the GT feces events were classified as *Small Urine* while 0% as *Large Urine*. No GT urine or GT feces event was classified as Large BG.

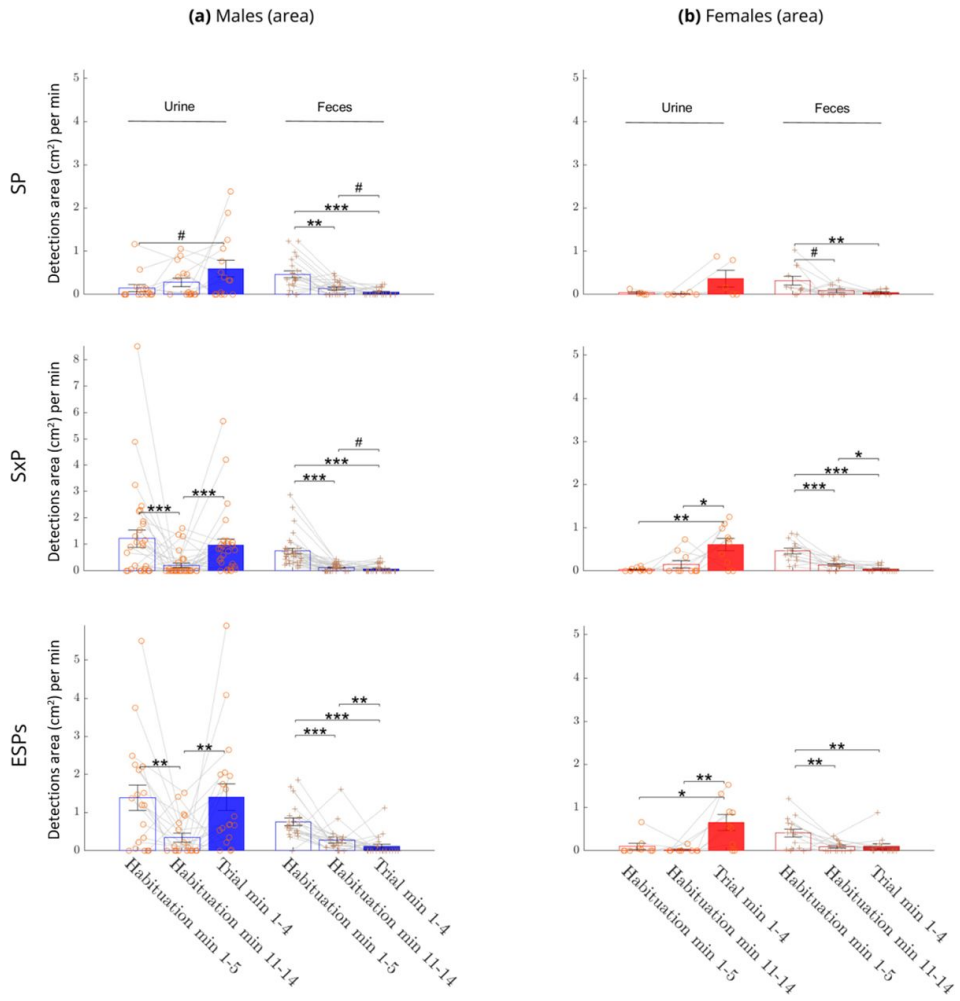


Figure 4-figure supplement 1.

Urine and fecal depositions area during habituation start, habituation end, and trial. The mean area ±SEM of urine and fecal depositions per minute during habituation start (minutes 1-5), habituation end (minutes 11-14), and trial (first four minutes of trial). Statistical comparisons between the three periods (three pair-wise comparisons) were done separately for urine and fecal depositions. Mice with no urine or feces detection in these periods were ignored from the urine or feces analysis, respectively.

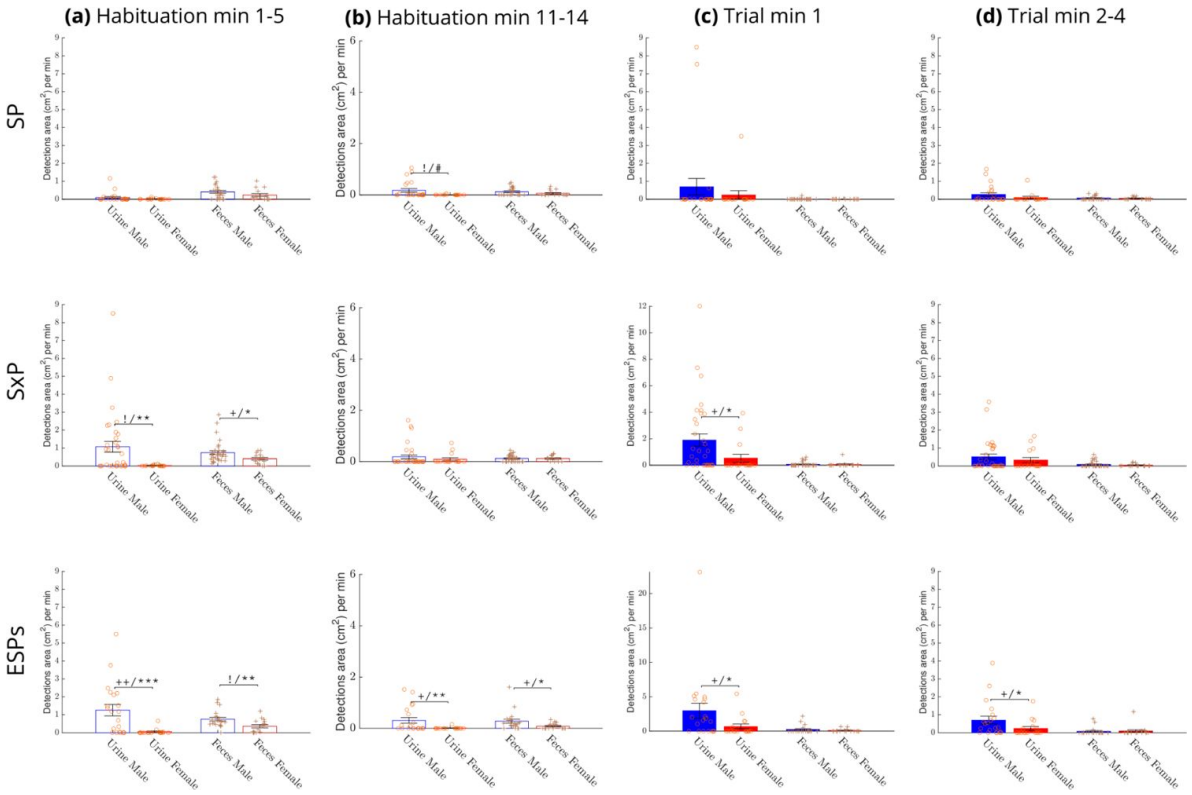


Figure 5—figure supplement 1.

Comparison of mean deposition areas between sexes. The mean area ±SEM of urine and fecal depositions in males (blue bars) vs. females (red bars) during early (minutes 1-5) and late (minutes 11-14) minutes of habituation and during the first minute and minutes 2-4 of the trial. A significant difference between the mean area of urine or fecal depositions (Wilcoxon rank sum test) is marked with * (or # for 0.05<p-value ≤0.1) and a significant difference in the distribution of non-depositing animals (Chi-square test) is marked with + (or ! for 0.05<p-value ≤0.1).

References

Arakawa H, Blanchard DC, Arakawa K, Dunlap C, Blanchard RJ (2008) **Scent marking behavior as an odorant communication in mice** *Neuroscience & Biobehavioral Reviews* **32**:1236–1248

Bigiani A, Mucignat-Caretta C, Montani G, Tirindelli R. (2005) **Pheromone reception in mammals** *Reviews of Physiology, Biochemistry and Pharmacology* :1–35 https://doi.org/10.1007/s10254-004-0038-0

Brennan PA (2004) **The nose knows who's who: chemosensory individuality and mate recognition in mice** *Hormones and Behavior* **46**:231–240 https://doi.org/10.1016/j.yhbeh.2004.01.010

Brennan PA, Kendrick KM (2006) **Mammalian social odours: attraction and individual recognition** *Philosophical Transactions of the Royal Society B: Biological Sciences* **361**:2061–2078

Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) **End-to-end object detection with transformers** :213–229

Dalghi MG, Montalbetti N, Wheeler TB, Apodaca G, Carattino MD (2023) **Real-time void spot assay** *JoVE (Journal of Visualized Experiments* **192**

Desjardins C, Maruniak J, Bronson F (1973) **Social rank in house mice: differentiation revealed by ultraviolet visualization of urinary marking patterns** *Science* **182**:939–941

He K, Zhang X, Ren S, Sun J (2016) **Deep residual learning for image recognition** :770–778

Higuchi Y, Arakawa H (2022) **Contrasting central and systemic effects of arginine-vasopressin on urinary marking behavior as a social signal in male mice** *Hormones and Behavior* **141**

Hurst JL, Beynon RJ (2004) **Scent wars: the chemobiology of competitive signalling in mice** *Bioessays* **26**:1288–1298

Hyun M, Taranda J, Radeljic G, Miner L, Wang W, Ochandarena N, Huang KW, Osten P, Sabatini BL (2021) **Social isolation uncovers a circuit underlying context-dependent territory-covering micturition** *Proceedings of the National Academy of Sciences* **118**

Keil KP, Abler LL, Altmann HM, Bushman W, Marker PC, Li L, Ricke WA, Bjorling DE, Vezina CM (2016) **Influence of animal husbandry practices on void spot assay outcomes in C57BL/6J male mice** *Neurourology and urodynamics* **35**:192–198

Keller JA, Chen J, Simpson S, Wang EHJ, Lilascharoen V, George O, Lim BK, Stowers L (2018) **Voluntary urination control by brainstem neurons that relax the urethral sphincter** *Nature neuroscience* **21**:1229–1238

Kopachev N, Netser S, Wagner S (2022) **Sex-dependent features of social behavior differ between distinct laboratory mouse strains and their mixed offspring** *Iscience* **25**

Mervis CB *et al.* (2012) **Duplication of GTF2I results in separation anxiety in mice and humans** *American Journal of Human Genetics* **90**:1064–1070 https://doi.org/10.1016/j.ajhg.2012.04.012

Miller CH, Haxhillari K, Hillock MF, Reichard TM, Sheehan MJ (2023) **Scent mark signal investment predicts fight dynamics in house mice** *Proceedings of the Royal Society B* **290**

Miller CH, Hillock MF, Yang J, Carlson-Clarke B, Haxhillari K, Lee AY, Warden MR, Sheehan MJ (2023) **Dynamic changes to signal allocation rules in response to variable social environments in house mice** *Communications Biology* **6**

Mohapatra AN, Peles D, Netser S, Wagner S (2024) **Synchronized LFP rhythmicity in the social brain reflects the context of social encounters** *Communications Biology* **7**

Netser S, Haskal S, Magalnik H, Bizer A, Wagner S (2019) **A system for tracking the dynamics of social preference behavior in small rodents** *JoVE (Journal of Visualized Experiments* **153**

Netser S, Haskal S, Magalnik H, Wagner S (2017) **A novel system for tracking social preference dynamics in mice reveals sex-and strain-specific characteristics** *Molecular autism* **8**:1–14

Verstegen AM, Tish MM, Szczepanik LP, Zeidel ML, Geerling JC (2020) **Micturition video thermography in awake, behaving mice** *Journal of neuroscience methods* **331**

Wegner KA *et al.* (2018) **Void spot assay procedural optimization and software for rapid and objective quantification of rodent voiding function, including overlapping urine spots** *American Journal of Physiology-Renal Physiology* **315**:F1067–F1080

Wöhr M, Roullet FI, Hung AY, Sheng M, Crawley JN (2011) **Communication impairments in mice lacking Shank1: reduced levels of ultrasonic vocalizations and scent marking behavior** *PloS one* **6**

Wolff P, Powell A (1984) **Urine patterns in mice: an analysis of male/female counter-marking** *Animal behaviour* **32**:1185–1191

Editors

Reviewer #1 (Public Review):

Summary:
The manuscript provides a novel method for the automated detection of scent marks from urine and feces in rodents. Given the importance of scent communication in these animals and their role as model organisms, this is a welcome tool.

Strengths:
The method uses a single video stream (thermal video) to allow for the distinction between urine and feces. It is automated.

Weaknesses:
The accuracy level shown is lower than may be practically useful for many studies. The accuracy of urine is 80%. This is understandable given the variability of urine in its deposition, but makes it challenging to know if the data is accurate. If the same kinds of mistakes are maintained across many conditions it may be reasonable to use the software (i.e., if everyone is under/over counted to the same extent). Differences in deposition on the scale of 20% would be challenging to be confident in with the current method, though differences of the magnitude may be of biological interest. Understanding how well the data maintain the same relative ranking of individuals across various timing and spatial deposition metrics may help provide further evidence for the utility of the method.

https://doi.org/10.7554/eLife.100739.1.sa3

Reviewer #2 (Public Review):

Summary:
The authors built a tool to extract the timing and location of mouse urine and fecal deposits in their laboratory set up. They indicate that they are happy with the results they achieved in this effort.

The authors note urine is thought to be an important piece of an animal's behavioral repertoire and communication toolkit so methods that make studying these dynamics easier would be impactful.

Strengths:
With the proposed method, the authors are able to detect 79% of the urine that is present and 84% of the feces that is present in a mostly automated way.

Weaknesses:
The method proposed has a large number of design choices across two detection steps that aren't investigated. I.e. do other design choices make the performance better, worse, or the same? Are these choices robust across a range of laboratory environments? How much better are the demonstrated results compared to a simple object detection pipeline (i.e. FasterRCNN or YOLO on the raw heat images)?

The method is implemented with a mix of MATLAB and Python.

One proposed reason why this method is better than a human annotator is that it "is not biased." While they may mean it isn't influenced by what the researcher wants to see, the model they present is still statistically biased since each object class has a different recall score. This wasn't investigated. In general there was little discussion of the quality of the model. Precision scores were not reported. Is a recall value of 78.6% good for the types of studies they and others want to carry out? What are the implications of using the resulting data in a study? How do these results compare to the data that would be generated by a "biased human?"

5 out of the 6 figures in the paper relate not to the method but to results from a study whose data was generated from the method. This makes a paper, which, based on the title, is about the method, much longer and more complicated than if it focused on the method. Also, even in the context of the experiments, there is no discussion of the implications of analyzing data that was generated from a method with precision and recall values of only 70-80%. Surely

this noise has an effect on how to correctly calculate p-values etc. Instead, the authors seem to proceed like the generated data is simply correct.

https://doi.org/10.7554/eLife.100739.1.sa2

Reviewer #3 (Public Review):

Summary:
The authors introduce a tool that employs thermal cameras to automatically detect urine and feces deposits in rodents. The detection process involves a heuristic to identify potential thermal regions of interest, followed by a transformer network-based classifier to differentiate between urine, feces, and background noise. The tool's effectiveness is demonstrated through experiments analyzing social preference, stress response, and temporal dynamics of deposits, revealing differences between male and female mice.

Strengths:
The method effectively automates the identification of deposits
The application of the tool in various behavioral tests demonstrates its robustness and versatility.
The results highlight notable differences in behavior between male and female mice

Weaknesses:
The definition of 'start' and 'end' periods for statistical analysis is arbitrary. A robustness check with varying time windows would strengthen the conclusions.
The paper could better address the generalizability of the tool to different experimental setups, environments, and potentially other species.
The results are based on tests of individual animals, and there is no discussion of how this method could be generalized to experiments tracking multiple animals simultaneously in the same arena (e.g., pair or collective behavior tests, where multiple animals may deposit urine or feces).

https://doi.org/10.7554/eLife.100739.1.sa1

Author response:

We want to thank the reviewers for their constructive feedback.

General

The recall values of our method range between 78.6% for all urine cases to 83.3% for feces (and not between 70-80%, as stated by reviewer #2), with a mean precision of 85.6%. This is rather similar to other machine learning-based methods commonly used for the analysis of complicated behavioral readouts. For example, in the paper presenting DeepSqueak for analysis of mouse ultrasonic vocalizations (Coffey et al. DeepSqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations. *Neuropsychopharmacol*. 44, 859–868 (2019). https://doi.org/10.1038/s41386-018-0303-6), the recall values reported for both DeepSqueak, Mupet and Ultravox (Fig. 2c, f) are very similar to our method.

We have analyzed and reported all the types of errors made by our methods, which are mostly technical. For example, depositions that overlap the mouse blob for too long till getting cold will be associated with the mouse and therefore will not be detected (“miss” events). These technical errors are not supposed to create a bias for a specific biological condition and, hence, shouldn’t interfere with the use of our method. A video showing all of the mistakes made by our algorithm on the test set was submitted (Figure 2-video 1).

Below we will to relate to specific points and describe our plan to revise the manuscript accordingly.

Detection accuracy

a. It should be noted that when large urine spots are considered, our algorithm got 100% correct classification (Figure 2, supplement 1, panel b). However, small urine deposits are very similar to feces in their appearance in the thermal picture. In fact, if the feces are not shifted, discrimination can be quite challenging even for human annotators. To demonstrate the accuracy of the proposed method relative to human annotators, we plan to compare its results with the accuracy of a second human annotator.

b. As part of the revision, we plan to test general machine learning-based object detectors such as faster-RCNN or YOLO (as suggested by Reviewer 2) and compare them with our method.

c. To check if our method may introduce bias to the results, we plan to check if the errors are distributed evenly across time, space, and genders.

Design choices

(A) The preliminary detection algorithm has several significant parameters. These are:

a. Minimal temperature rise for detection: 1.1°C rise during 5 sec.

b. Size limits of the detection: 2 - 900 pixels.

c. Minimal cooldown during 40 sec: 1.1°C and at least half the rise.

d. Minimal time between detections in the same location: 30 sec.

We chose to use low thresholds for the preliminary detection to allow detection of very small urinations and to minimize the number of “miss” events, relying on the classifier to robustly reject false alarms. Indeed, we achieved a low rate of miss events: 5 miss events for the entire test set (1 miss event per ~90 minutes of video). We attribute these 5 “miss” events to partial occlusion of the detection by the mouse.

To adjust the preliminary detection parameters to a new environment, one will need to calibrate these parameters in their own setup. Mainly, the size of the detection depends on the resolution of the video, and the cooldown rate might be affected by the material of the floor, as well as the room temperature.

We plan to explore the robustness of these parameters in our setup and report the influence on the accuracy of the preliminary algorithm.

(B) We chose to feed the classifier with 71 seconds of videos (11 seconds before the event and 60 seconds after it) as we wanted the classifier to be able to capture the moment of the deposition, the cooldown process, as well as urine smearing or feces shifting which might give an additional clue for the classification. In the revised paper we plan to report accuracy when using a shorter video for classification.

Generability

a. In the revised version, we plan to report the accuracy of the method used on a different strain of mice (C57), with a different arena color (white arena instead of black).

Statistics

a. In the revised paper, we will explain why we chose each time window for analysis. Also, we will report statistics for different time windows, as suggested by Reviewer 3.

b. Unlike reviewer #2, we don't think that the small difference in recall rate between urine and feces (78.6% vs. 83.3%, respectively) creates a bias between them. Moreover, we don't compare the urine rate to the feces rate.

c. In the revised manuscript we will explicitly report the precision scores, although they also appear in our manuscript in Fig. 2- Supplement 1b.

https://doi.org/10.7554/eLife.100739.1.sa0